The Role of Pretest Measures of the Outcome in Causal Inference

By

Yongnam Kim

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy
(Educational Psychology)

at the
UNIVERSITY OF WISCONSIN–MADISON
2018

Date of final oral examination: 8/7/2018

This dissertation is approved by the following members of the Final Oral Committee:

Peter M. Steiner, Associate Professor, Educational Psychology
Jee-Seon Kim, Professor, Educational Psychology
David Kaplan, Professor, Educational Psychology
James Wollack, Professor, Educational Psychology
Felix Elwert, Professor, Sociology

Dedicated to my parents

ACKNOWLEDGEMENTS

Many people around me have helped me in preparing and finishing this dissertation. I thank my advisor, Peter M. Steiner, for his help and guidance, especially for always motivating me through conversation. I thank Jee-Seon Kim and David Kaplan whose objective perspectives made them ideal readers of my work. I thank James Wollack for his advanced seminar class where I have felt comfortable to discuss my ideas. I thank Felix Elwert for encouraging me to investigate the topic presented in Study 3 of this dissertation. Regarding Study 3, I also thank M. H. Clark for providing access to her data sets for the illustration. Most importantly, I thank my wife, Sonyoung Yim, for her love and support. I thank my son, Danu, for reminding me that this dissertation is just one of the minor things I do in my life.

TABLES OF CONTENTS

LIST OF TABLES

< STUDY 2 >

< STUDY 3 >

LIST OF FIGURES

< STUDY 1 >

< STUDY 2 >

< STUDY 3 >

< APPENDIX B >

ABSTRACT

Among all types of covariates, pretest measures of the outcome have gained special attention in the educational and psychological literature. Despite the popularity and emphasis in research practice, however, the theoretical justification and the specific role of pretests in making a causal inference have not been explicitly discussed. This dissertation is a collection of three papers each of which focuses on the specific context of the identification of causal effects using pretests. Study 1, “Gain Scores Revisited: A Graphical Models Perspective,” investigates how gain scores, defined as the differences between posttest and pretest measures, can be used to adjust for confounding bias. One of the main findings is that gain score estimators are robust against the unreliability of pretests, bias amplification, and collider bias. Study 2, “Causal Graphical Views of Fixed Effects Models,” generalizes the identification results of Study 1 to fixed effects models, in which the gain score model is considered as one of several analytic models for analyzing pretest-posttest data under fixed effects models. Study 2 argues that, contrary to popular belief, the alleged problem of time-invariant covariates is not a limitation of the fixed effects approach. Study 3, “Causal Identification Using Difference-in-Differences in Mediation Analysis,” discusses how pretests can be used in causal mediation analysis to relax the stringent, no unmeasured mediator-outcome confounding assumption. Using real data sets, Study 3 shows that the proposed mediation approach can outperform the standard mediation approach in the presence of unmeasured mediator-outcome confounding.

INTRODUCTION

The fundamental difficulty in causal inference is that the treatment-outcome relationship can be explained not only by the causal relationship between the treatment and the outcome but also by alternative spurious associations between them. In order to rule out non-causal explanations and to uniquely identify causal effects, researchers frequently collect additional information about study units, referred to as *covariates*. A covariate is any variable, other than the treatment and the outcome, that describes units' characteristics and is recorded before starting the treatment (Rosenbaum, 2002). For example, when study units are individuals, subjects' gender, age, education level, or other factors that describe their characteristics can serve as covariates.

Among all types of covariates, pretest measures of the outcome, simply referred to as *pretests*, have been particularly emphasized by researchers in education and psychology (e.g., Campbell & Stanley, 1963; Cook & Steiner, 2010; Hallberg, Cook, Steiner, & Clark, 2016; Shadish, Cook, & Campbell, 2002; Wong, Valentine, & Miller-Bains, 2017). Shadish et al. (2002) wrote that for causal inference, “no single variable will usually do as well as the pretest” (p. 136). Pretest measures are one of the critical elements of quasi-experimental designs developed by Campbell and his colleagues (Campbell & Stanley, 1963; Shadish et al., 2002) and are found in many educational and psychological studies (Huck & McLean, 1975; Kenny, 1975; Weiss, 1972).

Despite this enduring emphasis and popularity, however, a formal theory of the role of pretests in causal inference has not been explicitly discussed in the literature. Many studies have mainly focused on empirical analysis, but the empirical results of the effect of pretests are somewhat unclear. Wong et al. (2017) reported in their review of 12 empirical studies that although the use of pretests often reduces bias, it does not always eliminate it. Also, Hallberg and her colleagues found that a higher correlation between the pretest and the posttest does not ensure more bias-reduction (Hallberg et al., 2016). St.Clair, Cook, and Hallberg (2014) even reported that the use of multiple pretests, compared to a single pretest, can increase bias. Without a theoretical explanation about the roles and the mechanisms of pretests, these empirical results are hard to interpret.

Also, many prior studies have almost exclusively focused on only one of the roles that pretests can play: bias-removing by conditioning on pretests. For example, individual cases are matched on pretests (matching), or pretests are controlled for in regression analysis (lagged dependent variable regression). However, “pretests serve many purposes” (Shadish et al., 2002, p. 136). Nonetheless, in the literature, it is unclear how and under which conditions pretests play different roles, other than being conditioned on, and how the use of pretests can improve the quality of causal investigations.

This dissertation is a collection of three papers, each of which formalizes the special role that pretest measures of the outcome play in different causal contexts. In the remainder of this dissertation, each of the three papers is presented in order. Study 1 discusses the gain score model, which is well-known in the educational and psychological literature. Importantly, Study 1 finds and formalizes a unique bias-removing mechanism using pretests, which is fundamentally different from the regular bias-removing mechanism in matching or regression analysis. The unique mechanism is referred to as the *bias-offsetting* mechanism. The findings are extended to Study 2 where the gain score model is considered as one of several analytic models for analyzing data under fixed effects models. The new formalization of fixed effects models helps to clarify prevalent misunderstandings about the fixed effects approach. Based on the findings of the bias-offsetting mechanism, Study 3 develops a novel method to identify causal mediation effects using pretests. No unmeasured mediator-outcome confounding is one of the fundamental problems in causal mediation analysis, and this problem can be resolved by relying on the bias-offsetting mechanisms with pretests. Each study can be viewed as a single complete manuscript such that it has its own introduction and discussion. The overarching conclusion of this dissertation is presented in the final section.

STUDY 1

GAIN SCORES REVISITED: A GRAPHICAL MODELS PERSPECTIVE

Abstract

The use of gain scores for identifying causal treatment effects has been frequently avoided among social scientists. This paper develops graphical models and graph-based arguments to show that gain score methods are a viable strategy to identify causal effects in observational studies. Our graphical models reveal that gain score methods use a different bias-removing mechanism than matching or covariance adjustments that aim at blocking non-causal association. Gain score methods offset, rather than block, the non-causal association relying on the common trend assumption. We show that due to the unique mechanism, gain score estimators are robust against measurement error in the pretest, bias amplification, and collider bias. We also clarify why assessing the common trend assumption becomes intractable when the pretest directly affects the treatment assignment. Finally, we discuss the distinct role of pretests in the context of Lord's paradox.

INTRODUCTION

Pretest or baseline measures of the outcome, or simply *pretests*, have gained much attention in the literature in the social sciences (e.g., Campbell & Stanley, 1963; Cook & Steiner, 2010; Shadish, Cook, & Campbell, 2002). The corresponding literature as well as empirical evidence from meta-analyses suggest that pretest measures are the most important covariates for removing confounding bias in observational studies (Cook, Shadish, & Wong, 2008; Hallberg, Cook, Steiner, & Clark, 2016; Wong, Valentine, & Miller-Bains, 2017). Such pretests can be used to compute and analyze *gain scores*, also called change or difference scores, which represent the differences between the posttest and pretest scores (Allison, 1990; Kenny, 1975; Maris, 1998). Nonetheless, gain score methods have long been criticized and frequently avoided by applied researchers and methodologists. Campbell and Erlebacher (1970, p. 197) wrote that “gain scores are in general such a treacherous quicksand,” and Cronbach and Ferby (1970, p. 80) even recommended researchers to “frame their questions in other ways.” This negative view is still widespread among researchers even until recently (Smolkowski, 2013; Thomas & Zumbo, 2012).

Instead, researchers have preferred covariance adjustments or matching methods that control for or match on the pretest (or a corresponding propensity score) in order to estimate the causal effect of an intervention (e.g., Imbens & Wooldridge, 2009). We refer to these methods as *conditioning methods* because the causal effects are obtained “conditional” on the pretest (and other covariates or the corresponding propensity score). Causal identification using conditioning methods relies on the *unconfoundedness* assumption, also called strong ignorability or conditional independence assumption (Imbens, 2004; Rosenbaum & Rubin, 1983). Meeting the unconfoundedness assumption requires that researchers know all the confounding covariates and measure them reliably (Steiner, Cook, & Shadish, 2011). Since this is rarely the case, the use of conditioning methods in observational studies frequently results in biased effect estimates.

We argue that gain score methods are a viable alternative to identify causal effects when the unconfoundedness assumption is violated. Although the causal assumption underlying gain score methods, the *common trend* assumption, might not be fully met either, gain score estimators have at least three advantages over conditioning estimators (e.g., matching or covariance adjustment estimators): They are immune to (i) unreliability of the pretest, (ii) bias amplification, and (iii) collider bias. Especially, gain score estimators' robustness to bias amplification and collider bias has never been discussed in the literature despite the longstanding discussions about gain scores, particularly in the context of Lord's paradox (Lord, 1967). As we will graphically show, these comparative advantages originate from the different bias-removing mechanisms of gain score and conditioning estimators. While conditioning methods remove bias via *blocking* non-causal association, gain score methods remove bias via *offsetting* non-causal association by differencing rather than conditioning.

We use graphical models to discuss the identification strategy of gain scores and their advantages in identifying causal effects. Graphical models are a visual representation of the structural causal model of the presumed data-generating process for the data at hand. This approach has been developed in computer sciences (Pearl, 1988) and epidemiology (Robins, 1987), and it is now becoming more popular also in the social sciences (e.g., Elwert, 2013; Morgan & Winship, 2015; Steiner, Kim, Hall, & Su, 2017). The use of graphical models has two major advantages over algebraic formulations in the discussion of causal identification. First, graphical models allow us to discuss causal assumptions and bias-removing mechanisms in an intuitively appealing but nonetheless formal rigorous way. With graphs and graph-based arguments, we can literally *see* the common trend assumption and the bias-offsetting mechanism of gain score methods. Second, graphical representations of subject-matter theory provide an indispensable tool in practice for assessing the common trend assumption's plausibility. This enables researchers to better defend (or reject) the rather abstract common trend assumption.

In this article, we consider a nonrandomized two-group pretest-posttest design, where the outcomes of the treatment and control group are measured at two points in time, before and after the intervention. Given the pretest measure, researchers have two main choices to identify and estimate the treatment effect. They can use *conditioning* methods like matching or covariance adjustments, or gain score methods—the classic setting of Lord's paradox (Lord, 1967). To ease exposition, we restrict our discussion to linear data-generating models with constant effects across all units (extensions to nonlinear or nonparametric settings is a topic for future research). In discussing gain score and conditioning estimators, we focus our attention exclusively on bias and do not discuss any efficiency or significance testing issue. This does not mean that efficiency and testing can be ignored in practice, but our major aim here is to guide researchers in choosing an identification and estimation strategy that results in the least possible bias.[1]

The article is organized as follows. In the next section, we provide a brief introduction to graphical models for observational studies and gain scores, and discuss the assumptions and mechanisms necessary for identifying causal treatment effects. In the following section, we highlight the three advantages of gain score estimators over conditioning estimators. The section is followed by an in-depth discussion of the common trend assumption. We discuss scenarios and conditions under which the assumption is not met or is hard to assess. We conclude with a discussion of the distinct role of pretests in observational studies.

[1] Van Breukelen (2006) discussed efficiency issues of gain score estimators.

Graphical Models Perspectives on Observational Studies and Gain Scores

Graphical Models for Observational Studies

It is well known that causal inference with observational data is challenging because the treatment and control groups are frequently not comparable at baseline (Shadish et al., 2002). In the absence of randomly assigned treatment and control conditions, the observed group difference in the outcome reflects not only potential causal effects but also spurious associations due to confounding (i.e., differential selection of units into the treatment and control groups). However, if researchers succeed to reliably measure a set of covariates that meets the unconfoundedness assumption, then the causal effect is identified and can be correctly estimated via matching or covariance adjustments (provided technical assumptions for matching and covariance adjustment, e.g., correct specification of the functional form).

The above rationale can be *visualized* by causal graphical models. They facilitate our intuitive understanding without sacrificing formal rigor. Consider an example where we are interested in evaluating the effect of participating in a summer math camp (Z) on students' math achievement (Y). Assume that participation in the math camp was not randomized, instead students or their parents decided whether to enroll or not. Further assume that we know, from subject-matter theory and empirical investigations, that students' true but latent math ability (A) is the sole confounding variable that affects both treatment Z and outcome Y. Figure 1.1A shows the corresponding graphical model consisting of three nodes and three arrows. The nodes represent the variables and the arrows the causal relationships between the nodes. For instance, the arrow $A \rightarrow Z$ indicates that students' math ability causally affects participation in the math camp (e.g., high ability students might more likely enroll the camp than low ability students). Since math ability A also affects math achievement Y ($A \rightarrow Y$), A is referred to as a confounding

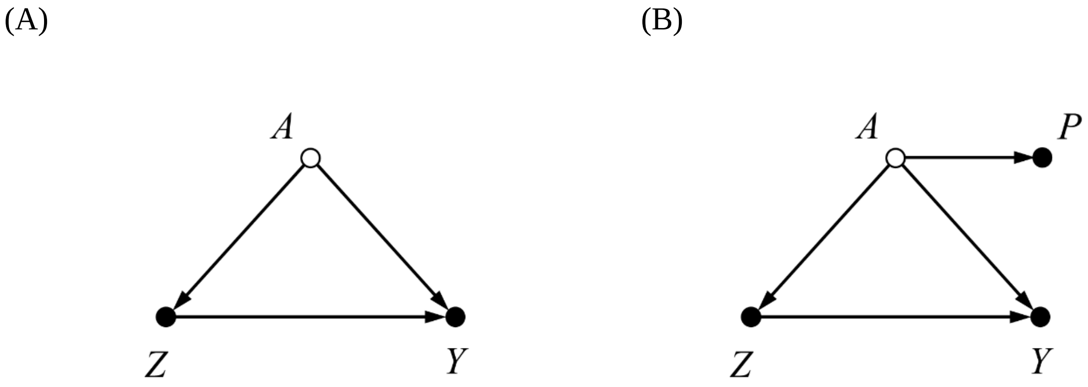

FIGURE 1.1. *Graphs for observational studies. (A) Graph for an observational study without pretest. (B) Graph for an observational study with pretest.*

variable or confounder because *A* confounds the relationship between treatment *Z* and outcome *Y*. Since *A* is unmeasured, its node is vacant; observed nodes are filled. It is important to note that the causal graph describes how the data were actually generated, regardless of whether a variable has been measured. Thus, a graphical model is a graphical representation of the presumed data-generating process, and it typically contains all observed but also unobserved variables that directly or indirectly affect both treatment and outcome.

Given the graphical model in Figure 1.1A, we see that treatment *Z* and outcome *Y* are connected or associated via two different *paths*[2]:

(i) $Z \rightarrow Y$;

(ii) $Z \leftarrow A \rightarrow Y$.

The first path represents the *causal* relationship of interest while the second path represents a *non-causal* relationship between *Z* and *Y*. Both paths are naturally open and thus transmit

[2] A path is a sequence of adjacent nodes without visiting a node more than once. The directions of the arrows do not matter.

association. The paths are “naturally” (i.e., without any other conditioning) open because they do not contain a *collider* (Elwert & Winship, 2014; Pearl, Glymour, & Jewell, 2016). A collider is a node at which two arrows from its adjacent nodes collide (e.g., C in $A \rightarrow C \leftarrow B$ is a collider variable). A path with a collider does not transmit association without any other conditioning because any association terminates at the collider node, that is, the path is naturally blocked. Therefore, the overall association between Z and Y is a mixture of the causal and non-causal associations. Unless the non-causal association via path (ii) is stripped out, the observed marginal association between Z and Y does not correspond to the causal relationship between Z and Y only via path (i).

The naturally open non-causal paths can be *blocked* by conditioning on any middle node in the paths unless it is a collider. Since A is the sole middle node and not a collider on path (ii), conditioning on A via matching or regression blocks the transmission of non-causal association. Conditional on A, the association between Z and Y is then only determined by the causal association transmitted via path (i), $Z \rightarrow Y$. Thus, the causal effect is identified conditional on A. Pearl (1993) developed a simple graphical criterion, the *backdoor criterion*, to test whether a set of observed variables is sufficient to identify causal effects via conditioning. The backdoor criterion states that causal effects are identified if all non-causal (or backdoor) paths can be blocked. For our graph in Figure 1.1A, however, the non-causal path $Z \leftarrow A \rightarrow Y$ cannot be blocked because the ability A is latent (a vacant node) and thus unavailable for conditioning. Thus, the causal effect of attending the math camp on math scores is not identifiable via matching or covariance adjustments.

Although the confounder A is unmeasured, researchers may have a pretest measure of the outcome that may serve as a proxy for A. For example, one may measure students’ math achievement before the math camp starts. Let *P* denote such a pretest measure. Then, both pretest and posttest are likely affected by students' math ability. Figure 1.1B shows the graph with the added pretest: *A* affects both *P* and *Y*, but *Z* does not affect *P* (because *P* is measured before *Z*). Since *P* is measured (filled node), we can condition on it (e.g., matching on *P* or regressing *Y* on *Z* and *P*). However, conditioning on *P* does still not identify the causal effect because *P* is not a middle node on the non-causal path $Z \leftarrow A \rightarrow Y$ and thus cannot block the path. Due to the pretest's correlation with ability, conditioning on *P* may reduce the confounding bias but it cannot eliminate all the bias. As the graphical model in Figure 1.1B demonstrates, conditioning on a pretest measure is hardly sufficient to identify causal effects in observational studies.

Graphical Models for Gain Scores

In the presence of unmeasured confounding, gain score methods can be an alternative strategy to identify causal effects. Gain score methods first require computing the gain score: $G = Y - P$. In Figure 1.2A, the gain score *G* is added as a new node to the graph. Since *G* is determined by both *Y* and *P*, we add two arrows: $P \rightarrow G$ and $Y \rightarrow G$. Moreover, since the gain score is computed as a linear combination of *P* and *Y* with fixed coefficients of -1 and $+1$, respectively, we also add the corresponding structural coefficients to the graph in Figure 1.2A (see Pearl, 2016, for a similar graphical representation of gain scores; also, Shahar & Shahar, 2012). Assuming linear relationships and constant effects, we now label all arrows with Greek letters, which represent the unknown structural coefficients of the underlying data-generating process. For example, τ on the causal path $Z \rightarrow Y$ represents the constant causal effect of *Z* on *Y*.

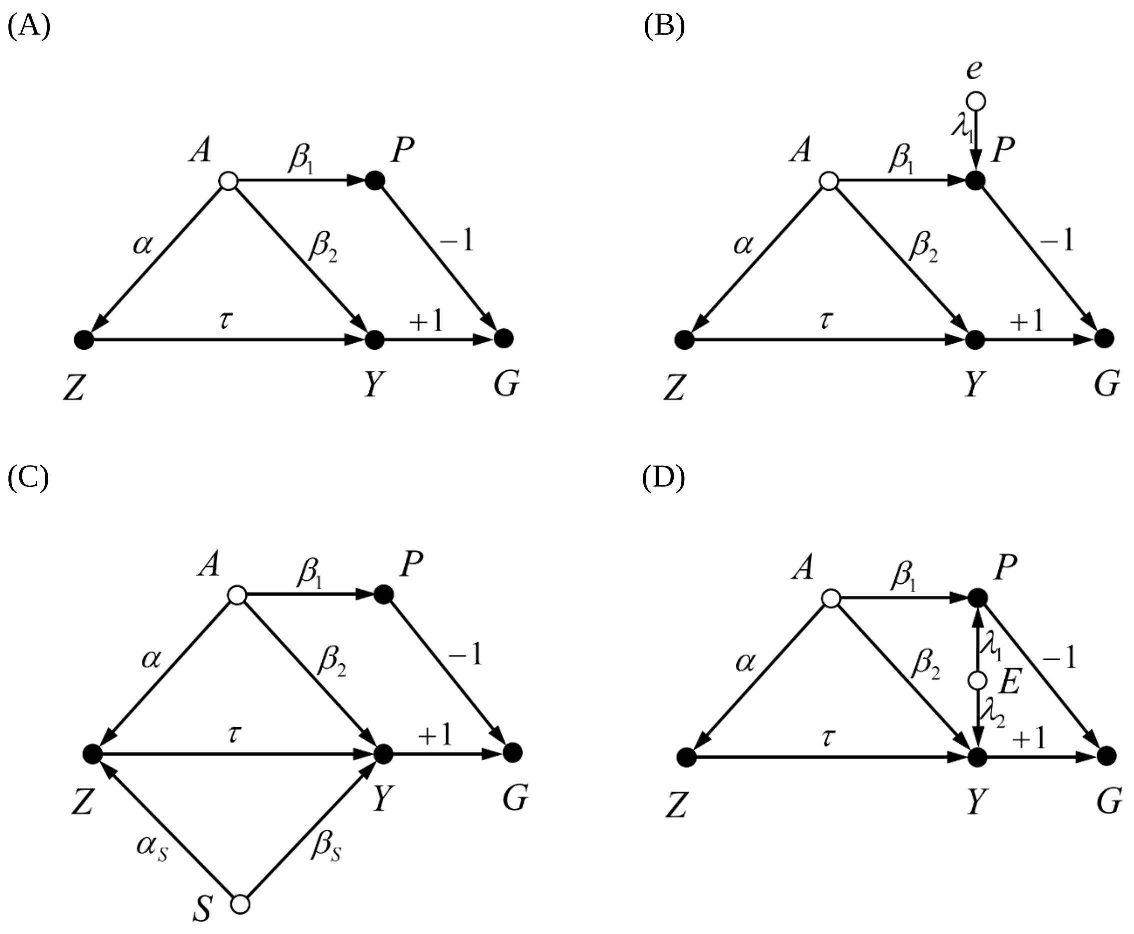

FIGURE 1.2. *Graphs for gain scores. (A) Simple gain score graph. (B) Graph showing how the independent measurement error e affects the pretest. (C) Graph with two confounders where the pretest is only affected by confounder A. (D) Graph with a common measurement error that affects both the pretest and posttest.*

Gain score methods investigate the causal effect of Z on the gain score G rather than the original outcome Y. That is, instead of Y, we regress G on Z or, equivalently, compare the group mean difference in the gain score G using a two-sample t-test. Since the causal effect of Z on G

is the mediated effect via the causal path $Z \to Y \to G$, the effect is given by the product of the two path coefficients: $\tau \times (+1) = \tau$, which is identical to the causal effect of Z on Y.[3]

However, with regard to the causal relationship between Z and G in Figure 1.2A, we now have three non-causal paths:

(i) $Z \leftarrow A \to P \to G$;

(ii) $Z \leftarrow A \to Y \to G$;

(iii) $Z \to Y \leftarrow A \to P \to G$.

Since the non-causal paths (i) and (ii) are naturally *open* (they do not contain any collider variable), they transmit associations and confound the relation between Z and G. In contrast, path (iii) is naturally *blocked* by the collider Y and thus does not transmit any association. According to the backdoor criterion, the causal effect of Z on G is identified when the two open non-causal paths (i) and (ii) are blocked. Although conditioning on both P and Y would block the non-causal paths (note that A is unmeasured),[4] conditioning on Y would also block the causal path $Z \to Y \to G$. Since the causal path must remain unblocked, conditioning on both P and Y is not a viable identification strategy.

Gain score methods eliminate the confounding bias by *offsetting* rather than *blocking* the non-causal associations. The association transmitted via the non-causal path (i) can be quantified

[3] The path-tracing computation of effects was developed by Wright (1921). For more information about Wright's path-tracing rules, see Pearl (2013).

[4] In fact, conditioning on Y opens the path (iii) because Y is a collider on the path. But, additional conditioning on P blocks the path and thus the path (iii) does not transmit any association, conditional on Y and P.

by the product of the three structural path coefficients on the path: $\alpha \times \beta_1 \times (-1) = -\alpha\beta_1$.[5] Analogously, we can quantify the associations via the other non-causal paths:

(i) $Z \leftarrow A \rightarrow P \rightarrow G,\ -\alpha\beta_1;$

(ii) $Z \leftarrow A \rightarrow Y \rightarrow G,\ +\alpha\beta_2;$

(iii) $Z \rightarrow Y \leftarrow A \rightarrow P \rightarrow G,\ 0.$

Note that the path (iii) transmits no association because this path is naturally blocked by the collider *Y*. If the sum of all the non-causal associations is zero,

$$\alpha\beta_2 - \alpha\beta_1 = \alpha(\beta_2 - \beta_1) = 0,$$

the non-causal associations offset each other, and all the confounding bias is eliminated. Because α is assumed to be nonzero,[6] the confounding bias cancels out if the unmeasured math ability *A* affects the pretest math score *P* and the posttest math score *Y to the same extent*, $\beta_1 = \beta_2$. This equality condition is frequently referred to as the *common trend* (Lechner, 2011) or *time-invariant confounding* assumption. If the common trend assumption holds, the relation between the treatment and the gain score is free of any confounding and the overall association between *Z* and *G* is solely due to the causal effect of *Z* on *G*. Importantly, the bias-removing mechanism of offsetting confounding bias does not require any conditioning to block non-causal paths. Therefore, the causal effect can be identified by gain score methods despite the presence of unmeasured confounders (i.e., violation of the unconfoundedness).

[5] To be precise, the product should be multiplied by the variance of the “root” node of the path, such that $-\alpha\beta_1 \times Var(A)$. Throughout this paper, we assume that unmeasured confounders (i.e., vacant nodes) such as *A*, have a unit-variance, $Var(A) = 1$.

[6] If $\alpha = 0$, we would have no arrow *A* → *Z*, implying that *A* is not a confounder.

It is interesting to investigate what happens if we were to condition on the pretest *P* in a gain score analysis, for instance, regressing *G* on *Z* and *P*. As the gain score graph directly reveals, conditioning on *P* blocks the non-causal path (i) $Z \leftarrow A \rightarrow P \rightarrow G$ while the non-causal path (ii) $Z \leftarrow A \rightarrow Y \rightarrow G$ remains open. This results in losing the bias offsetting effect of gain score methods and we have the same bias as in a standard matching or covariance adjustment with respect to *Y* (Allison, 1990; Jamieson, 2004; Kenny, 1975; Laird, 1983; Lechner, 2011). This is so because the association transmitted via the remaining open path (ii) $Z \leftarrow A \rightarrow Y \rightarrow G$ is identical to the non-causal association via $Z \leftarrow A \rightarrow Y$. Thus, conditioning on the pretest in a gain score analysis does not use the unique bias-removing mechanism of the gain score methods but turns the analysis into standard conditioning methods.

ADVANTAGES OF GAIN SCORE ESTIMATORS

Unreliability of Pretest

In practice, pretests are often contaminated with random measurement error. To highlight the impact of measurement error on both conditioning and gain score estimators, we now explicitly add an independent error term *e* to the graph in Figure 1.2B (by convention, such random disturbance terms are usually omitted from graphs). The new structural parameter λ_1 represents the impact of the measurement error *e* on the pretest *P* ($e \rightarrow P$).

Given the graph, conditioning on *P* will eliminate more bias if *P* closely resembles *A*, which is the case whenever measurement error is very small. However, as measurement error *e* increases, *P* becomes a poorer proxy for *A*, resulting in more bias in conditioning estimators. This intuition is confirmed if we consider the regression of *Y* on *Z* and *P* and express the expectation of *Z*'s partial regression coefficient, $b_{YZ|P}$, in terms of the structural parameters in the graph in Figure 1.2B (for derivations of all estimator formulae in Study 1, see Appendix A):

$$b_{YZ|P} = \tau + \frac{\alpha\beta_2(1-r)}{Var(Z)-\alpha^2 r}, \tag{1.1}$$

where *r* denotes the reliability of *P*, $r = \frac{\beta_1^2}{\beta_1^2+\lambda_1^2}$.[7] The regression estimator consists of the true causal effect (τ) and an additive bias term. The bias term shows that the regression coefficient varies with the impact of the measurement error, λ_1. For example, if measurement error is large (i.e., $|\lambda_1|$ is large in comparison to $|\beta_1|$), the reliability *r* decreases and the bias term $|\alpha\beta_2(1-r)|$ increases. That is, $(1-r)$ % of the confounding bias induced by *A* (i.e., $\alpha\beta_2$) is remaining. In addition, the remaining bias is amplified by the factor $\frac{1}{Var(Z)-\alpha^2 r}$ (we discuss bias amplification in the next section). If the pretest is measured without error (i.e., $\lambda_1 = 0$ and $r = 1$), then the partial regression estimator is unbiased, $b_{YZ|P} = \tau$. Thus, measurement error attenuates the bias-removing potential of the pretest (see Aiken & West, 1991; Steiner et al., 2011).

In contrast to conditioning estimators, gain score estimators are insensitive to measurement error in the pretest (Maris, 1998). This is so because the association transmitted via the non-causal path $Z \leftarrow A \rightarrow P \rightarrow G$ does not involve λ_1. According to the path-tracing rule, the association along the path is simply given by the product of the three path coefficients of α, β_1,

[7] Note that we assume $Var(A) = Var(e) = 1$.

and -1. In regressing the gain score *G* on the treatment indicator *Z*, we can write the expectation of the gain score estimator, b_{GZ}, in terms of structural parameters τ, α, β_1, and β_2:

$$b_{GZ} = \tau + \frac{\alpha(\beta_2 - \beta_1)}{Var(Z)}. \tag{1.2}$$

The formula clearly shows that the gain score estimator is not a function of λ_1 (or the reliability *r*), revealing its insensitivity to measurement error in the pretest.[8] Suppose that the math pretest is a highly unreliable measure of students' true math ability. In this case, conditioning methods are not able to remove all the bias. Depending on the unreliability, only a minor fraction of the bias might be removed. However, gain score methods' accuracy is unaffected by the measurement error. As long as the common trend assumption holds, $\beta_1 = \beta_2$, they identify the causal effect without any bias.

Bias Amplification

Steiner and Kim (2016) showed that any remaining bias in conditioning estimators is *amplified* (also see Pearl, 2010, 2011). Bias amplification is a phenomenon that occurs with conditioning methods whenever the conditioning covariates (a) fail to remove the entire bias *and* (b) causally determine treatment selection. The denominator in the bias term of Equation (1.1) contains the amplification factor $\frac{1}{Var(Z) - \alpha^2 r}$, which is always greater than the factor without conditioning on *P*, $\frac{1}{Var(Z)}$. Thus, bias left due to the pretest's unreliability is strongly amplified because of conditioning on *P*, which results in the subtraction of $\alpha^2 r$ in the denominator.

[8] However, the gain score estimator's variance will be affected by the unreliability in *P*.

However, this does not occur with gain scores estimators in Equation (1.2). The denominator of the bias term does not have the subtraction of $\alpha^2 r$.

To better see this, consider two unobserved confounders *A* and *S* as depicted by the graph in Figure 1.2C. The new variable *S* is a confounder of the causal relation between *Z* and *Y* because *S* simultaneously affects the treatment and outcome. Since the relations *S* → *Z* and *S* → *Y* are described by the structural parameters α_s and β_s , respectively, the confounding bias induced by *S* is given by $\alpha_s \beta_s$. The graph also shows that the pretest *P* is affected by *A* while unaffected by *S*, indicating that *P* can serve as a proxy for *A* but not for *S*. Here, conditioning on *P* does not eliminate any bias induced by the confounder *S*, on the contrary, it amplifies the bias due to *S*.

Given the graph in Figure 1.2C, the expected partial regression coefficient $b_{YZ|P}$ (i.e., conditioning estimator) is

$$b_{YZ|P} = \tau + \frac{\alpha\beta_2(1-r)}{Var(Z)-\alpha^2 r} + \frac{\alpha_S \beta_S}{Var(Z)-\alpha^2 r} . \tag{1.3}$$

The first bias term represents, as already discussed, the remaining bias due to *P*'s unreliability with respect to *A*, while the second bias term shows the hidden bias due to *S*, $\alpha_s \beta_s$, which is amplified by the factor $\frac{1}{Var(Z)-\alpha^2 r}$. In order to see that bias amplification only occurs if we condition on the pretest *P*, compare Equation (1.3) to the expected regression estimator without conditioning on *P* (i.e., regression of *Y* on *Z*):

$$b_{YZ} = \tau + \frac{\alpha\beta_2}{Var(Z)} + \frac{\alpha_S \beta_S}{Var(Z)} . \tag{1.4}$$

It becomes clear that, without conditioning on *P*, the hidden biases due to *A* and *S* are not amplified because $\alpha^2 r$ is not subtracted from *Var*(*Z*) in the denominators.

Simply speaking, since bias amplification is a phenomenon that only occurs if one *conditions* on covariates, gain score estimators are immune to bias amplification (provided one does not condition on any other covariates[9]). Regressing *G* on *Z*, the expectation of the gain score estimator is given by

$$b_{GZ} = \tau + \frac{\alpha(\beta_2 - \beta_1)}{Var(Z)} + \frac{\alpha_S \beta_S}{Var(Z)} . \qquad (1.5)$$

Note that the third term, the bias due to *S* (i.e., $\alpha_S \beta_S / Var(Z)$), is identical to the third term in Equation (1.4), which is the bias in the unadjusted effect estimate of *Z* on *Y*. Therefore, although gain score methods do not eliminate the bias due to *S*, at least they do not amplify the bias as in the use of conditioning methods.

Collider Bias

Since pretest and posttest are typically measured with the same or a very similar instrument (e.g., same or same type of test or questionnaire items, or interviewers) in the same or a similar setting (e.g., lab or classroom, lab personnel or teachers), the error terms of the pretest and posttest are very likely correlated. Zimmerman and Williams (1982) wrote, “correlated errors [between pretests and posttests] are probably the *rule* rather than the *exception* in pretest-posttest measurements” (p. 153, emphasis added). The graph in Figure 1.2D represents such a

[9] It is possible that one conditions on covariates (other than pretests) in a gain score analysis. This may be desirable because the common trend assumption can be met only after conditioning on some covariates. Although this strategy may introduce bias amplification in gain score estimators, in this paper we consider a basic form of gain score estimators, which does not require any other conditioning.

correlated error structure. The exogenous variable E represents a common source of the correlated measurement errors, that is, E simultaneously affects the pretest and posttest with structural parameters λ_1 and λ_2 for the causal relations $E \to P$ and $E \to Y$, respectively.

Given the data-generating model in Figure 1.2D, conditioning methods now face the issue of *collider bias*. Compared to the graph in Figure 1.2A, where error terms are independent, the correlated error structure in Figure 1.2D creates an additional non-causal path between Z and Y:

$$Z \leftarrow A \to P \leftarrow E \to Y.$$

Since this path contains the collider P, it is naturally blocked at the collider node P. However, once we condition on P, the path becomes unblocked and starts transmitting spurious association (Ding & Miratrix, 2015; Elwert & Winship, 2014). This spurious association between Z and Y is referred to as the *collider bias*. Thus, with correlated errors, conditioning estimators are biased due the unreliable measurement of A but also the collider bias due to conditioning on the collider P.

When it comes to gain score methods, the graph in Figure 1.2D reveals that the correlated error structure via E creates three additional non-causal paths between Z and G:

(i) $Z \leftarrow A \to P \leftarrow E \to Y \to G$;

(ii) $Z \leftarrow A \to Y \leftarrow E \to P \to G$;

(iii) $Z \to Y \leftarrow E \to P \to G$.

However, gain score methods are robust against collider bias because all new non-causal paths via E are naturally blocked either at P or Y because one of them is always a collider on the path. Thus, no non-causal association is transmitted through these three non-causal paths. Since gain score methods condition neither on *P* nor on *Y*, the non-causal paths remain naturally blocked such that collider bias is not an issue for gain score estimators.

This can also be seen from algebraic expressions of the conditioning and gain score estimators. According to the data-generating model in Figure 1.2D, the expectation of the conditioning estimator is given by

$$b_{YZ|P} = \tau + \frac{\alpha\beta_2(1-r)}{Var(Z)-\alpha^2 r} - \frac{\alpha\beta_1\lambda_1\lambda_2}{\{Var(Z)-\alpha^2 r\}Var(P)} . \qquad (1.6)$$

Compared to Equation (1.1), the correlated error structure results in an additional subtractive bias term—the collider bias. This new bias term corresponds to the unblocked collider path *Z* ← *A* → *P* ← *E* → *Y*, given by the product of the four structural path coefficients of the path: α , β_1, λ_1, and λ_2.

In comparison to conditioning methods, gain score methods are unaffected by collider bias. The expectation of the gain score estimator for the graph in Figure 1.2D is indeed identical to Equation (1.2). Therefore, although a common cause *E* is responsible for the correlated error structure, such a dependence does not affect the gain score estimators.

The Common Trend Assumption under Extended Data-Generating Models

When Pretest Affects Treatment

The common trend assumption means that time trends of the repeated measures (i.e., pretest and posttest) do not differ between the treatment and control groups in the absence of the treatment effect. Given the previous data-generating models (except for Figures 1.1A and 1.2C), the common trend assumption is equivalent to the requirement that the impacts of the unmeasured confounder *A* on *P* and *Y* are identical, $\beta_1 = \beta_2$. However, whether or not the equality establishes the common trend assumptions strongly depends on the actual data-generating model. The models thus far assumed that the pretest *P* has no causal effect on treatment *Z* and posttest *Y*. This might often be unrealistic in practice. For example, the math pretest score may be known to students and their parents before they decide whether to attend the math camp or not. Then, parents of students with a low pretest score may encourage their children to take the camp, that is, the pretest causally affects the treatment selection ($P \rightarrow Z$).

The graph in Figure 1.3A describes this scenario. The graph has four naturally open non-causal paths between *Z* and *G* with the following transmitted associations:

(i) $Z \leftarrow A \rightarrow P \rightarrow G,\ -\alpha\beta_1$;

(ii) $Z \leftarrow A \rightarrow Y \rightarrow G,\ +\alpha\beta_2$;

(iii) $Z \leftarrow P \rightarrow G,\ -\gamma_1 Var(P)$;

(iv) $Z \leftarrow P \leftarrow A \rightarrow Y \rightarrow G,\ +\gamma_1\beta_1\beta_2$.

Note that *P* is a “root” node of path (iii). To obtain an unbiased estimate of the causal effect using gain score methods, the sum of the four non-causal associations must be zero,

$$\alpha(\beta_2 - \beta_1) + \gamma_1\{\beta_1\beta_2 - Var(P)\} = 0 .$$

One obvious case that meets this condition is $\beta_1 = \beta_2$ and $\beta_1\beta_2 = Var(P)$. While the former equality, $\beta_1 = \beta_2$, is easy to conceptualize (i.e., math ability affects both pretest and posttest math achievements to the same extent), the substantive meaning of the latter equality, $\beta_1\beta_2 = Var(P)$, is hard to understand. Algebraically, if the variance of the pretest’s error term is $Var(e) = -\beta_1(\beta_2 - \beta_1)$, then $\beta_1\beta_2 = Var(P)$ is met, but we do not see any subject-matter

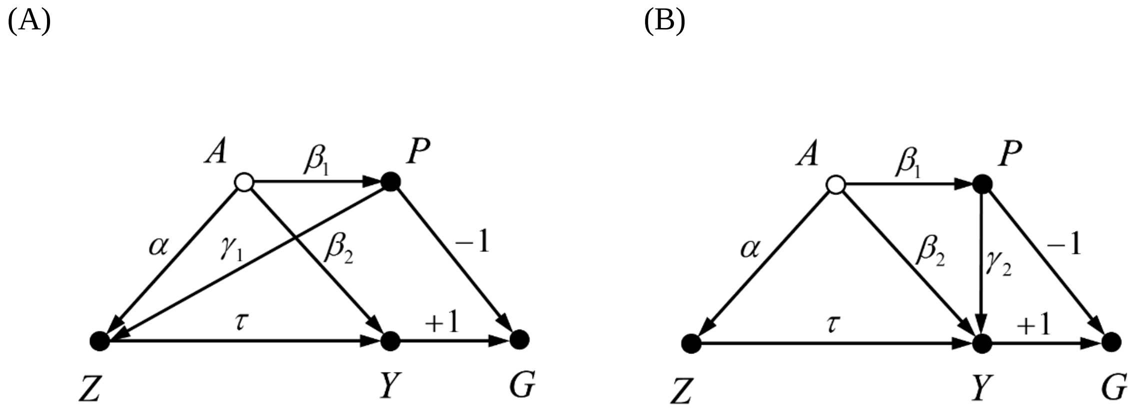

FIGURE 1.3. *Graphs where the pretest directly affects (A) the treatment or (B) the posttest.*

rationale why the pretest's error term should adhere to this restriction.[10] Thus, given the data-generating model in Figure 1.3A, it is generally hard to assess whether the common trend assumption for gain score methods actually holds.

It is worth noting that the association via the non-causal path (iii) above depends on the variance of the pretest, $Var(P)$, which implies that measurement error in the pretest now affects the gain score estimator. Thus, the robustness of gain score methods to the unreliability of the pretest no longer holds when the pretest measure directly affects the treatment as in Figure 1.3A.

When Pretest Affects Posttest

The graph in Figure 1.3B describes a situation where the pretest causally affects the posttest ($P \rightarrow Y$). Such a situation occurs if the pretest scores are unknown prior to camp enrollment, but after students and parents learn about the pretest score, it may stimulate students'

[10] One interpretable case is when the pretest is perfectly measured without any measurement error. Given $Var(P) = Var(\beta_1 A + e)$, where *e* is the random measurement error of *P*, $\beta_1\beta_2 = Var(P)$ is identical to $\beta_1(\beta_2 - \beta_1) + Var(e) = 0$. Thus, given $\beta_1 = \beta_2$, $Var(e) = 0$ satisfies the equation. However, this perfect measurement rarely happens in practice.

motivation or parents' engagement to organize private tutoring, for instance. With respect to the gain score *G*, we have three naturally open non-causal paths and corresponding associations:

(i) $Z \leftarrow A \rightarrow P \rightarrow G,\ -\alpha\beta_1;$

(ii) $Z \leftarrow A \rightarrow Y \rightarrow G,\ +\alpha\beta_2;$

(iii) $Z \leftarrow A \rightarrow P \rightarrow Y \rightarrow G,\ +\alpha\beta_1\gamma_2.$

In order to identify the causal effect, the common trend assumption requires that

$$\alpha\{\beta_2 + \beta_1(\gamma_2 - 1)\} = 0.$$

Given $\alpha \neq 0$, the equality holds when $\beta_1 = \beta_2 + \beta_1\gamma_2$. Note that $\beta_2 + \beta_1\gamma_2$ represents the total effect of *A* on *Y* (except for the effect via *Z*): the direct effect of *A* on *Y* (*A* → *Y*) plus the mediated effect (*A* → *P* → *Y*). Thus, the common trend assumption requires that the impact of *A* on *P* is the same as the total impact of *A* on *Y* (again, not via *Z*). This requirement is indeed interpretable.

Generally, gain score methods are not recommended when the pretest directly affects the treatment or the posttest (Allison, 1990; Maris, 1998; also see Imai & Kim, 2017, for a similar advice for fixed effects models). Here, we demonstrated the difference between the two situations. When the pretest directly affects the treatment (*P* → *Z*), the common trend assumption is hardly justifiable, and gain score estimators become sensitive to the unreliability of the pretest. In contrast, when the pretest directly affects the posttest (*P* → *Y*), the common trend assumption might still be defendable and gain score estimators remain unaffected by the unreliability in *P*.

Discussion

The widespread reservations about gain scores are partly due to the lack of understanding about how gain score methods actually remove bias. A few methodological articles argued that gain score methods can be effective for causal inference with observational studies (e.g., Allison, 1990; Maris, 1998; Van Breukelen, 2006). However, most of these articles rely exclusively on algebra, which is not easily accessible to many applied researchers. This paper revisited the topic with a graphical models approach. Our graphical representations visualize the process of how gain score methods identify causal effects. We imposed causal contexts and interpretations on the common trend assumption such that researchers can more easily communicate and evaluate the assumption in their studies. For example, we clarified why assessing the common trend assumption becomes intractable when the pretest directly affects the treatment assignment. We also showed that gain score estimators are robust against the unreliability of pretests, bias amplification and collider bias—issues that may strongly affect conditioning estimators.

Of course, gain score methods do not always work and not necessarily remove more bias than conditioning methods like regression or matching adjustments. It is possible that conditioning on the pretest yields less biased or even unbiased effect estimates while gain score estimators might be seriously biased. The message of this paper is that researchers need subject-matter knowledge about the data-generating process to select an appropriate method. Without strong subject-matter knowledge, an informed choice of an identification strategy and corresponding estimator is impossible. This is why this difficulty has been called a “paradox,” as in Lord’s (1967) paradox. We do not even know which method might perform better “on average” in practice. However, if some subject-matter knowledge is available, graphical models are a useful tool to incorporate such knowledge into causal investigations. The graphs then help us in assessing which identification strategy has the best chances to work and why other strategies might not work. As they help in clarifying missing data problems (Thoemmes & Mohan, 2015) and quasi-experimental designs (Steiner et al., 2017), graphical models also help in understanding gain score methods and Lord's paradox.

This paper suggests a distinct role of pretests in observational studies. Although the literature has emphasized the importance of pretests (Cook & Steiner, 2010; Shadish et al., 2002), it has been unclear what a good pretest is. Our comparison of gain score and conditioning estimators revealed that different methods exploit different characteristics of the pretests. For conditioning estimators, a good pretest serves as good *proxy* of the unmeasured confounders, that is, the pretest and the unobserved confounders should be *nearly perfectly associated*. For gain score estimators, however, a good pretest is affected by the unobserved confounders *to nearly the same extent* as the posttest. When planning an observational study, researchers need to assess whether it is easier to closely meet the unconfoundedness or the common trend assumption with the pretest measure in their studies. If they think that the pretest may be a good proxy for all unobserved confounders, then they need to put considerable efforts into the reliable measurement of a single or multiple pretests and no correlated errors between the repeated measures. In contrast, if researchers believe that the unobserved confounders affect the pretest and posttest to almost the same extent, a reliable measurement and independent error of the pretest are less important. Instead, researchers might put more effort into using the same instrument at the pretest and posttest, or into an adequate calibrating or equating of scores from different instruments.

Finally, the graphical discussion of gain scores in this paper is a first step in developing graphical models for a broader class of methods including fixed effects models and comparative interrupted time series designs. They belong to the same class of methods as gain score methods because they rely on the bias offsetting mechanism of differencing instead of the bias-blocking mechanism of conditioning methods like matching or covariance adjustments. This distinction (conditioning methods vs. differencing methods) has not been clearly made in the previous literature. For example, it is not rare to find a comparative interrupted time series design with additional controlling for or matching on multiple pretests (e.g., St.Clair, Cook, & Hallberg, 2014; Wong, Valentine, & Miller-Bains, 2017; also see Abadie, Diamond, & Hainmueller, 2010, for synthetic control methods). Our graphs show that conditioning on the pretest when using a gain score, the regression model automatically turns into standard covariance adjustment, which then relies on the unconfoundedness rather than the common trend assumption. More research is needed to reveal similarities but also differences among those methods.

STUDY 2

CAUSAL GRAPHICAL VIEWS OF FIXED EFFECTS MODELS

Abstract

Fixed effects models, which have been developed in econometrics, are less well-known to applied psychological researchers. Using linear structural equation modeling, this paper introduces causal graphical views of fixed effects models and explains how they can be used to identify causal effects in observational studies. To formalize fixed effects models, we distinguish between data-generating models and analytic models. Identification of causal effects from data generated under fixed effects models can be achieved through several analytic models, including the gain score model, the deviation score model, or the dummy variable regression model. Our graphical representations, which merge data-generating models and analytic models, visualize the causal identification assumption of the fixed effects approach and the bias-removing mechanisms. Based on the proposed graphs and graph-based arguments, we also discuss some prevalent misunderstandings about the fixed effects approach. We argue that, although the alleged inability to control for time-invariant covariates might be a limitation of the dummy variable regression model, it is definitely not a limitation of the fixed effects approach itself. Also, by developing graphical representations for the random effects approach, we show why the approach necessarily results in biased effect estimates in observational studies.

INTRODUCTION

Finding causation from data is the key of psychological sciences. Traditionally, psychologists have preferred *experimentation* as a method for causal inference (Darley, 2001). Looking at the history of psychology, one finds many creative ideas that allow psychologists to investigate complex and invisible psychological phenomena in well-controlled lab settings (e.g., Leon Festinger's experiment regarding cognitive dissonance; Festinger & Carlsmith, 1959). Although such controlled experiments improve *internal validity* of a study, they may decrease *external validity* because a real-life setting is almost always different from a controlled lab setting (Cook & Campbell, 1979). A group of psychologists has developed and advanced a methodological framework called *quasi-experimentation* that can complement traditional experimentation (Shadish, Cook, & Campbell, 2002). The proposed quasi-experimental designs and methods provide psychologists with a methodological approach for investigating causation from non-experimental or observational data. Researchers from fields outside psychology, such as economics, sociology, epidemiology, and statistics, have also developed their own methods for causal inference, especially using observational data. Some of the methods overlap (e.g., matching, regression discontinuity), but there are some methods that are not well-known but are potentially useful in psychological studies.

Fixed effects models are one such method. This approach is very popular in other social sciences including economics, political science, and sociology, especially in studies using longitudinal or panel data, but it has been less used in psychology. The fixed effects approach has gained increasing attention because it allows for the identification of causal effects even in the presence of unmeasured confounding (Allison, 2009; Gunasekara, Richardson, Carter, & Blakely, 2014). This is attractive because in observational studies, it is very likely that the measured covariates do not capture all confounding. As such, one may consider using the fixed effects approach as an alternative to matching or any similar approaches which identify causal effects by conditioning on, or controlling for, the measured covariates. Of course, the fixed effects approach is not a silver bullet. It requires its own identification assumption and relies on unique bias-removing mechanisms. As the assumption and the mechanisms substantially differ from those of usual quasi-experimental designs and methods in psychology, it is important to understand how the fixed effects approach works so as to properly apply the approach to real psychological questions.

Unfortunately, the current literature on fixed effects models is not easily accessible to applied researchers, especially in psychology. This is because it is mostly in econometrics or statistics literature where discussions about fixed effects models occur. Introductory econometrics textbooks discuss fixed effects models, but they rely on complex algebraic expressions, following their convention; also, causal frameworks (e.g., potential outcomes or causal graphical models) have not been explicitly employed in many textbooks (e.g., Hsiao, 2003; Wooldridge, 2010, 2012). Recently, Sobel (2012) and Imai and Kim (2017) have explained how the fixed effects approach can be used to identify causal effects, but their intended audience is statisticians or econometricians, and their papers are formalized in a mathematically intensive way. Currently, fixed effects models remain alien to many applied psychologists.

The purpose of this paper is to introduce fixed effects models to applied psychologists using a language with which many of them are familiar. We explain the key causal identification features of the fixed effects approach using linear structural equation modeling (SEM). The linear SEM is one of the most popular research methods among social scientists, including psychologists, that has been used to describe the *causal* and *statistical* relationships among variables (Bollen & Pearl, 2013; Rohrer, 2018). By substituting the conventional algebraic formulation with the corresponding causal graphs (consisting of nodes and arrows) based on the linear SEM, we provide intuitive explanations about how the fixed effects approach can identify causal effects in observational studies. More precisely, our graphical representations, merging data-generating models and analytic models, show how each of the analytic models, such as the gain score model and the deviation score model, differently eliminates the unmeasured confounding bias that exists in the data under fixed effects models. In accordance with the distinction between data-generating and analytic models, hereafter we selectively use the term "fixed effects models" to indicate the data-generating models, and we use the term "the fixed effects approach" to indicate the analytic models or methods used for obtaining causal effects from data.

In addition, based on the proposed graphs and graph-based arguments, we revisit and clarify some prevalent misunderstandings or potentially misleading knowledge about the fixed effects approach. We explain that the alleged problem of dealing with time-invariant covariates (e.g., Baltagi, 2001; Clark & Linzer, 2015; Wooldridge, 2010, 2012) may be a limitation of the dummy variable regression model, which is just one of the analytic models, but other analytic models can easily control for such covariates and correct bias. We also develop graphical representations for the *random effects* approach and show that, because of its unique analytic feature (i.e., imperfect bias-offsetting), the approach necessarily results in biased effect estimates in observational studies.

We begin the rest of this paper with a brief introduction to the conventional algebraic formalization of fixed effects models. This formalization, which is common within the econometrics literature, then will be translated into graphical models. As we will show, graphical models are a powerful tool to formalize causal identification. In this paper, our concern is how the fixed effects approach identifies causal effects from data under fixed effects models, and we do not discuss any issues regarding standard errors of estimators (i.e., efficiency).

CONVENTIONAL ALGEBRAIC FORMALIZATION OF FIXED EFFECTS MODELS

Throughout this paper, we restrict our discussion to a simple pretest-posttest design or two-period panel data where the outcomes are measured at two points in time. Also, we consider that no one is exposed to the treatment in the first period. The key features of fixed effects models and the approach can be well understood using this simple design and data. But, the principles we present can be extended to more general cases. For example, see Appendix B for a case where the treatment regime is flexible enough that subjects may receive the treatment at any period of time.

Data-Generating Model

A *fixed effect* is a unit-specific effect that affects the outcome and that does not change over time or across cohorts. Although, in principle, it is possible that a fixed effect is invariant across cohorts at a particular point in time (e.g., Ashenfelter & Krueger, 1994; they used *twins* to deal with *family*-fixed effects; also see Turkheimer & Harden, 2014), a fixed effect is typically found with a *time* dimension as is exhibited in longitudinal or panel data. In many econometric textbooks, the basic form of fixed effects models where covariates are not considered is expressed as follows (e.g., Wooldridge, 2010, 2012):

$$Y_{it} = \tau A_{it} + \theta_i + \varepsilon_{it}, \tag{2.1}$$

where *i* denotes the unit of the analysis (e.g., individuals) and *t* denotes the time index, $t \in \{1 = before,\ 2 = after\}$. Among the variables, the outcome Y_{it}, the treatment A_{it}, and the idiosyncratic error (i.e., random noise) ε_{it} are all time-varying because they take the time index *t*. For example, Y_{i1} represents unit *i*'s outcome at $t = 1$, and ε_{i2} represents the idiosyncratic error the unit experiences at $t = 2$. In contrast, θ_i does not take the time index, and this is the unit-specific fixed effect with respect to time. The fixed effect is assumed to be correlated with the treatment, $Cov(A_{it}, \theta_i) \neq 0$, which is a typical assumption in observational studies. Otherwise, if they are uncorrelated with each other, it means that the treatment *A* is unconfounded as in randomized controlled trials; in our later discussion regarding random effects models, we consider this case further.

The meaning of Equation (2.1) is easy to understand when we split it by each time index. In a typical pretest-posttest design where no one is exposed to the treatment at $t = 1$, Equation (2.1) is expressed as

$$Y_{i1} = \tau A_{i1} + \theta_i + \varepsilon_{i1} = \theta_i + \varepsilon_{i1}, \qquad (2.2)$$

because $A_{i1} = 0$ for every unit *i*. Hereafter, for simplicity of notation, we may drop the unit subscript *i* if no confusion is likely. Equation (2.2) means that the pre-intervention measure Y_1, which is often referred to as the *pretest* (Shadish, Cook, & Campbell, 2002), is determined by (i) the unit-specific fixed effect θ and (ii) the idiosyncratic error ε_1 occurred at $t = 1$. Similarly, when $t = 2$, Equation (2.1) is given by

$$Y_{i2} = \tau A_{i2} + \theta_i + \varepsilon_{i2}. \qquad (2.3)$$

Equation (2.3) means that the post-intervention measure Y_2 is determined by (i) the same fixed effect θ, (ii) the idiosyncratic error ε_2 occurred at $t = 2$, and additionally, (iii) the effect of the treatment A_2 implemented at $t = 2$. The impact of the treatment A_2 on the post-intervention measure Y_2, which is τ, is the causal effect we are interested in.

The Problem of Causal Identification Under Fixed Effects Models

Identification of causal effects from the data generated by fixed effects models, described in Equations (2.1), (2.2), and (2.3), is rather challenging. Usual regression or matching methods fail to identify causal effects. Let's consider a naïve method that simply regresses the post-intervention measure Y_2 on the treatment indicator A_2. Fitting the regression model of $\hat{Y}_{i2} = a + bA_{i2}$, where *a* and *b* are, respectively, the intercept and the slope of the regression model, the population regression coefficient for A_2 is given by

$$\begin{aligned} b &= \frac{Cov(Y_2, A_2)}{Var(A_2)} = \frac{Cov(\tau A_2 + \theta + \varepsilon_2, A_2)}{Var(A_2)} \\ &= \tau + \frac{Cov(\theta, A_2)}{Var(A_2)}. \end{aligned} \tag{2.4}$$

The second bias term does not disappear because, as we discussed, the covariance or correlation between the fixed effect and the treatment is not generally zero in observational studies. Thus, we see that the naïve effect estimate of A_2 on Y_2 is biased.

Alternatively, one may consider using the pretest Y_1 as a control variable in the regression model. This is sometimes called a lagged dependent variable approach (Wooldridge, 2010). The lagged regression model has the form of $\hat{Y}_{i2} = a + bA_{i2} + cY_{i1}$ and aims to estimate τ

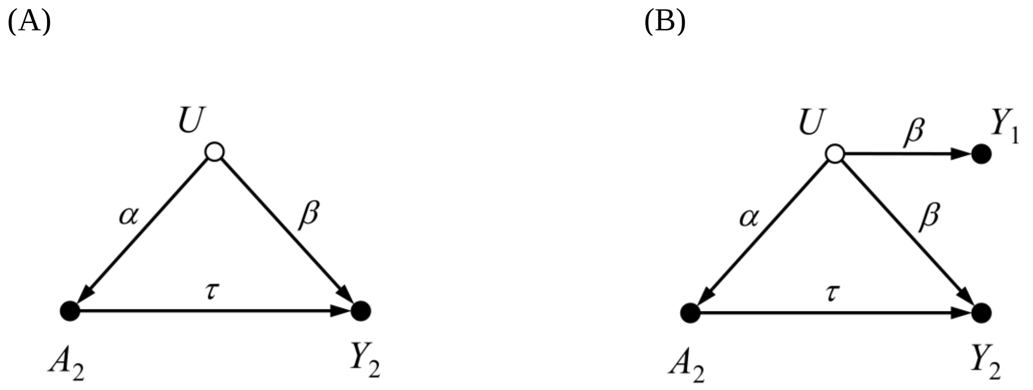

FIGURE 2.1. *Graphical representations for fixed effects models. (A) Variables occurred at $t = 2$ are represented. (B) Pre-intervention measure Y_1 is added to graph (A).*

by the partial regression coefficient *b* after controlling for Y_1. Unfortunately, in general, this does not produce unbiased effect estimates, either. The population partial regression coefficient of A_2 in the lagged regression model is given by

$$b = \tau + \frac{Cov(\theta, A_2)\left\{Var(Y_1) - Var(\theta)\right\}}{Var(A_2)Var(Y_1) - Cov(\theta, A_2)^2}. \tag{2.5}$$

See Appendix C for the derivation of the formula. As $Cov(\theta, A_2) \neq 0$ and $Var(Y_1) \neq Var(\theta)$, the bias term in Equation (2.5) does not vanish even though the pretest Y_1 is controlled for. Note that the variance equality holds only when $Var(\varepsilon_1) = 0$, implying no idiosyncratic or random error at $t = 1$. This is very unlikely to happen in practice.

GRAPHICAL REPRESENTATIONS FOR FIXED EFFECTS MODELS

Data-Generating Model

Although the literature on fixed effects models almost exclusively focuses on the algebraic formulation of data-generating models as exhibited in the previous section, *graphical models* or *directed acyclic graphs* (DAGs) in the linear SEM framework can also represent fixed effects models (e.g., Bollen & Brand, 2010; also see Imai & Kim, 2017). In Figure 2.1A, we present a graph that describes how the treatment and the outcome, at $t = 2$, are generated in observational studies. The graph represents that (i) the variable A_2 is causally determined by an unmeasured (indicated by vacant node) *U* and an omitted idiosyncratic error, and (ii) the variable Y_2 is causally determined by both *U* and A_2 and by its own, also omitted, idiosyncratic error.[1] Using the path coefficients on the arrows, the linear structural model depicting the causal process that generates Y_2 is given by

$$Y_{i2} = \tau A_{i2} + \beta U_i + \varepsilon_{i2}, \tag{2.6}$$

where ε_2 is the omitted idiosyncratic error at $t = 2$. Let's compare Equation (2.6) with Equation (2.3), $Y_{i2} = \tau A_{i2} + \theta_i + \varepsilon_{i2}$. The fixed effect θ is equivalent to the product of the parameter β and the unmeasured confounding factor *U*, $\theta_i = \beta U_i$. Thus, the fixed effect in the econometrics literature can be viewed as a combined entity of an unmeasured confounding factor and its impact on the outcome. This is the key insight in developing our graphical representations for fixed effects models. Imai and Kim (2016) suggested different graphical representations for fixed effects models by allowing for a separate node for the fixed effect θ. In the rest of this paper, we

[1] In graphical representations, idiosyncratic or random errors are typically omitted.

will show that our graphical representations provide a more intuitive understanding about the fixed effects approach.

If the pretest Y_1 is generated with the same fixed effect ($\theta_i = \beta U_i$) but with a different idiosyncratic error at $t = 1$, the linear structural model for it is given by

$$Y_{i1} = \beta U_i + \varepsilon_{i1}, \tag{2.7}$$

which corresponds to Equation (2.2). In Figure 2.1B, we describe this causal process of Y_1 by merging it with the previous graph in Figure 2.1A. The implication of fixed effects models is clear from the merged graph in Figure 2.1B. The repeated measures Y_1 and Y_2 are affected by the same confounder (*U*) to the same extent (β). We propose this graph as a graphical model that describes the data-generating process of fixed effects models, corresponding to Equation (2.1).

For our later discussion, it is worthwhile to further discuss two characteristics of the data-generating model depicted in Figure 2.1B. First, the model can be understood as a constrained or restricted model of a more general observational data model. One may imagine a model where the impacts of *U* on Y_1 and Y_2 are different. But, then this model would not belong to fixed effects models because there is no fixed effect (with respect to time) in this model; the effect is changing. Thus, the data-generating process by Equation (2.1) and Figure 2.1B has a stringent constraint or assumption regarding the equality of the impacts of *U* on the repeated measures Y_1 and Y_2. This has been referred to as the *common* (or *parallel*) *trend assumption* in the difference-in-differences literature (Lechner, 2011) and as the *fixed effects assumption* in the fixed effects models literature (Bell & Jones, 2015).

Second, such a strong constraint or assumption notwithstanding, usual methods such as the regular regression or matching methods still cannot identify the causal effect τ from the fixed effects model in Figure 2.1B. We already discussed this challenge by algebraically deriving Equations (2.4) and (2.5) where the bias terms do not vanish. But this result is directly found by using a simple and intuitive graphical criterion called the *backdoor criterion* (Pearl, 2009; Pearl, Glymour, & Jewell, 2016). In Figure 2.1B, the backdoor or non-causal path $A_2 \leftarrow U \rightarrow Y_2$ remains open i) when no variables are conditioned on, which corresponds to Equation (2.4), and ii) even when Y_1 is conditioned on, which corresponding to Equation (2.5). Unless all non-causal paths between the treatment and the outcome are blocked, the causal effect cannot be identified (see Rohrer, 2018, for an accessible introduction to graphical models and rules; she also provides many examples from psychological research).

Analytic Models for Fixed Effects Models

Despite the open backdoor path $A_2 \leftarrow U \rightarrow Y_2$ in Figure 2.1B, one may identify the causal effect τ if special analytic models or methods are applied to the data. In the econometrics literature, three methods are suggested: i) first differencing, ii) time demeaning, and iii) dummy variable regression (Wooldridge, 2010, 2012). The key features of those methods can be intuitively understood with graphical representations. As we will show below, the bias-removing mechanism of those analytic models creates another non-causal path(s) and offsets the existing non-causal path instead of blocking it.

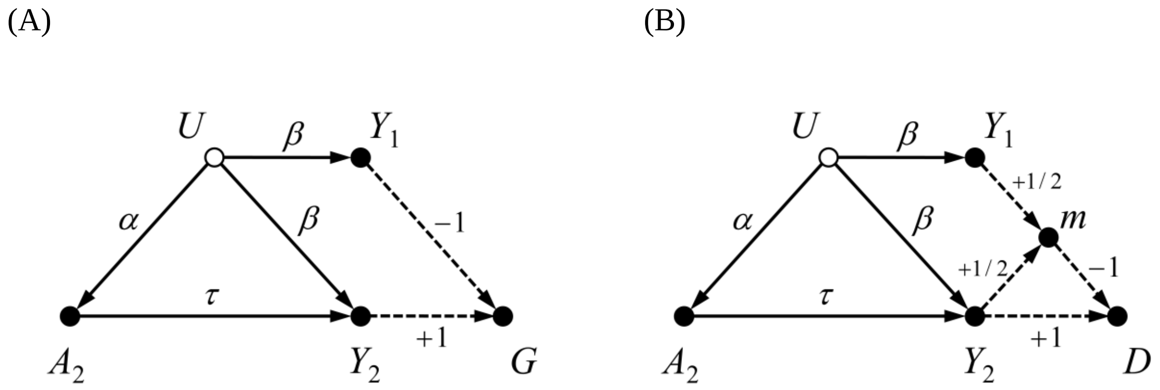

FIGURE 2.2. *Graphical representations, merging fixed effects data-generating models (Figure 2.1B) and analytic models. A gain score model (A) and a deviation score model (B) are added and represented by dashed arrows.*

Gain Score Model

We shall refer to the *first differencing* or *first difference model* in the econometrics literature as the *gain score model*, which would be more familiar to many psychologists. In Figure 2.2A, we describe the analytic model of the gain score model. In this graph, dashed arrows are used to distinguish the gain score model from the data-generating model which is represented by solid arrows. We use this arrow representation system to emphasize the fact that data-generating models cannot be altered by researchers (because they are generated by nature) whereas analytic models can be altered by researchers (because they are generated by researchers). Only appropriate analytic models that have been thoughtfully selected by researchers with substantial knowledge about the data-generating model allow us to identify causal effects from observational data.

When applying the gain score model, one first computes the gain score *G* by subtracting Y_1 from Y_2:

$$G_i = Y_{i2} - Y_{i1}. \tag{2.8}$$

The added structure using dashed arrows in Figure 2.2A represents Equation (2.8). As this variable *G* is created by researchers, the path coefficients on the dashed arrows $Y_1 \to G$ and $Y_2 \to G$ are exactly known as -1 and $+1$, respectively. In contrast, the path coefficients on the solid arrows remain unspecified, that is, we do not know what the values would be for the parameters α, β, and τ. However, the gain score model helps in identifying the unknown causal parameter τ. This identification procedure using graphs is suggested by Study 1 of this dissertation. Once *G* is computed, one investigates the effect of A_2 on the gain score *G*, not the original outcome Y_2, for example, by regressing *G* on A_2. In Figure 2.2A, the following three paths remain open between A_2 and *G*:

i) $A_2 \to Y_2 \to G$;

ii) $A_2 \leftarrow U \to Y_1 \to G$;

iii) $A_2 \leftarrow U \to Y_2 \to G$.

By the path-tracing rules (Wright, 1921; also see Pearl, 2013), the association transmitted via each path is easily computed. The product of the path coefficients along the path is multiplied by the variance of the “root” variable of the path. A root variable is a variable in the path that does not have any incoming arrows. For example, the root variable in path (i) is A_2, and the root variable in both path (ii) and path (iii) is *U*. Thus, each association via each of the paths above is respectively given by

i) $\tau \times (+1) \times Var(A_2) = Var(A_2)\tau$;

ii) $\alpha \times \beta \times (-1) \times Var(U) = -Var(U)\alpha\beta$;

iii) $\alpha \times \beta \times (+1) \times Var(U) = Var(U)\alpha\beta$.

The total association or the covariance between A_2 and *G* is the sum of the three partial associations. It is simply $Var(A_2)\tau$ because the second and the third associations offset each other, $Var(U)\alpha\beta - Var(U)\alpha\beta = 0$. From the graph, this can be understood that another non-causal path, path (ii), is created using the pretest Y_1, and this new path offsets the original non-causal path, path (iii). Fitting the regression model of $\hat{G}_i = a + bA_{i2}$, the population regression coefficient of A_2 is given by

$$b = \frac{Cov(A_2, G)}{Var(A_2)} = \frac{Var(A_2)\tau}{Var(A_2)} = \tau \ . \tag{2.9}$$

Thus, the causal effect τ is identified by the gain score model. For more detailed explanations of the gain score model, including its advantages over the ANCOVA model, see Study 1 of this dissertation.

Deviation Score Model

The gain score model, which is popular in the psychological literature is just one of the analytic models that researchers can choose for analyzing the data under fixed effects models. An alternative that is less known in the psychological literature but is very popular in the econometrics literature (Wooldridge, 2010, 2012) is the *time demeaning*, which we shall refer to as the *deviation score model*, following Allison (1994). True to its name, this model uses a deviation score instead of a gain score. A deviation score is defined as the difference between the posttest and the average of the repeated measures. In a simple pretest-posttest design, where only one pretest exists, we first compute the average *m* of the pretest Y_1 and posttest Y_2 for each unit *i*:

$$m_i = \frac{Y_{i1} + Y_{i2}}{2}. \tag{2.10}$$

Then, we compute the deviation score *D*:

$$D_i = Y_{i2} - m_i. \tag{2.11}$$

Finally, we investigate the treatment effect on the deviation score.

Figure 2.2B visualizes how the deviation score model can be added (dashed arrows) and be used to eliminate the unmeasured confounding bias from the same data-generating model in Figure 2.2A. First, as in Equation (2.10), the average node *m* is depicted as a common outcome of the pretest Y_1 and the outcome Y_2, $Y_1 \rightarrow m \leftarrow Y_2$. Again, we know the exact functional form of the mean computation, so the path coefficients on the two arrows, $Y_1 \rightarrow m$ and $Y_2 \rightarrow m$, are $+{}^1\!/_2$. The deviation score node *D* is determined from the outcome Y_2 and the average *m*, such that we have the corresponding arrows pointing into *D*, $Y_2 \rightarrow D \leftarrow m$. Again, from Equation (2.11), we know that the path coefficients of $Y_2 \rightarrow D$ and $m \rightarrow D$ are $+1$ and -1, respectively. Excepting causal paths between A_2 and *D*, all the open non-causal paths, together with the corresponding transmitted associations, are given by

i) $A_2 \leftarrow U \rightarrow Y_2 \rightarrow D,\ Var(U)\alpha\beta;$

ii) $A_2 \leftarrow U \rightarrow Y_2 \rightarrow m \rightarrow D,\ \ -Var(U)\dfrac{\alpha\beta}{2}$;

iii) $A_2 \leftarrow U \rightarrow Y_1 \rightarrow m \rightarrow D,\ \ -Var(U)\dfrac{\alpha\beta}{2}$.

Thus, the sum of all the three non-causal associations becomes zero, and the effect of A_2 on *D* is no longer confounded by *U* (or correctly adjusted for the bias due to *U*).

However, unlike in the gain score model, the effect of A_2 on *D* does not exactly correspond to the effect of A_2 on Y_2. In Figure 2.2B, the causal effect of A_2 on *D* is transmitted via the following two causal paths:

i) $A_2 \rightarrow Y_2 \rightarrow D,\ \ Var(A_2)\tau$;

ii) $A_2 \rightarrow Y_2 \rightarrow m \rightarrow D,\ \ -\dfrac{Var(A_2)\tau}{2}$.

The sum of the two associations is then $\dfrac{Var(A_2)\tau}{2}$, which is *half* of the causal association between A_2 on *G* in the gain score model, $Var(A_2)\tau$. Thus, fitting the regression model of $\hat{D}_i = a + bA_{i2}$, the population regression coefficient of A_2 is given by

$$b = \frac{Cov(A_2, D)}{Var(A_2)} = \frac{\tau}{2}. \tag{2.12}$$

Therefore, in order to recover the original causal quantity, the total effect of A_2 on *D* needs to be multiplied by *two*.[2] In fact, this is why many econometrics textbooks explain that independent

[2] Although we do not discuss it in this paper, if the repeated outcomes are measured at three points in time such that one may have one more prior measure Y_0 before Y_1 and Y_2, then the

variables also need to be time demeaned, just like the repeated outcome measures (Wooldridge, 2010, 2012). We see that such time demeaning procedures are required for different purposes. The time demeaning of the pretest and posttest outcome measures is necessary to offset the confounding bias while the time demeaning of independent variables, especially the treatment variable, is necessary to recover the original causal quantity.

Dummy Variable Regression

A computationally identical way of using the deviation score model is to use *dummy variable regression* (Allison, 1994; Lockwood & McCaffrey, 2007; Wooldridge, 2010, 2012). Instead of manually computing the deviation score, one may include individual dummy variables, indicating each *i*, into a pooled regression model such as,

$$\hat{Y}_{it} = bA_{it} + c_i ID_i + dT_t , \tag{2.13}$$

where ID_i is the set of individual dummy variables and T_t is the time indicator (i.e., $T_1 = 0$, $T_2 = 1$). Note that the intercept *a* in the regression model is absorbed into the dummies as a reference. Then, the regression coefficient *b* is numerically identical to the deviation score estimator, that is, the regression coefficient in Equation (2.12) × 2. Because of this equivalence, the deviation score estimator is sometimes referred to as the dummy variable estimator (Wooldridge, 2010). In our opinion, however, it is beneficial to distinguish between these two

computation slightly changes. First, the average *m* is computed by $m_i = (Y_{i0} + Y_{i1} + Y_{i2}) / 3$, and then the deviation score estimator from the regression of *D* (i.e., $D_i = Y_{i2} - m_i$) on A_2 should be multiplied by $3/2$. Drawing graphs and checking out all causal paths will easily verify this procedure.

analytic models or methods because they differ greatly in dealing with time-invariant covariates.[3] We will discuss this issue in more detail in the next section. Unfortunately, it is hard to graphically represent how the dummy variable regression model eliminates the unmeasured confounding bias, at least in an intuitively understandable way such as through the gain score and deviation score models depicted in Figure 2.2. So, we do not provide any graphical representation for the dummy variable regression model. One might consider that the dummy variable regression model *internally* creates the deviation score structure as depicted in Figure 2.2B, but the use of the individual dummies results in an important difference from the deviation score model.

ISSUES ON THE FIXED EFFECTS APPROACH

Controlling for Time-Invariant Covariates

A well-known alleged limitation of the fixed effects approach is that it cannot include, and thus cannot properly control for, time-invariant covariates (Allison, 2009; Bell & Jones, 2015; Clark & Linzer, 2015; Greene, 2011; Gunasekara et al., 2014; Halaby, 2004; Hsiao, 2003; Plümper & Troeger, 2007). For example, Wooldridge (2010) claims that "[w]hen analyzing individuals, factors such as gender or race cannot be included in [the fixed effects approach]" (p. 301). This alleged limitation is often contrasted with the fact that the random effects approach can include time-invariant covariates and is used as a justification for preferring the random effects approach over the fixed effects approach (e.g., Huber & Stephens, 2000; Nielsen & Alderson, 1995). In contrast, researchers who advocate for the fixed effects approach claim that

[3] Another difference is that they exploit different degrees of freedom. In the literature, some researchers claim that the use of dummy variable regression is desirable because it produces the correct standard errors (e.g., Allison, 2009; Wooldridge, 2010, 2012).

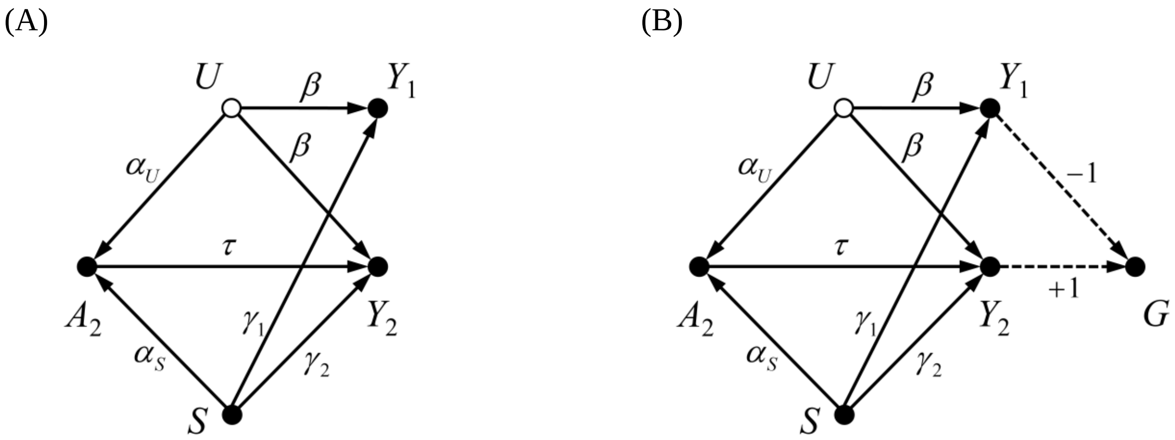

FIGURE 2.3. *Graphical representations with time-invariant covariate S. (A) Data-generating model. (B) Analytic model of the gain score model with the data-generating model (A).*

if the interest is not directly on the time-invariant covariates (because they are just control variables), the inability to estimate coefficients for the covariates is not a problem (Allison, 2009; Wooldridge, 2010). In what follows, we argue that the inability to control for time-invariant covariates is definitely not a limitation of the fixed effects approach. The inability might be a limitation of dummy variable regression, which is one of the analytic models for the fixed effects approach, but other analytic models (e.g., the gain score or deviation score model) can easily include and control for time-invariant covariates. Furthermore, we provide simple methods to control for time-invariant covariates in the dummy variable regression framework.

Time-Invariant Covariates Whose Impacts Are Time-Varying

Let us first consider how we could describe a time-invariant covariate such as gender or race within the graphical representations we have thus far considered. In Figure 2.3A, we add a covariate *S*, which is independent of *U*. Although *U* remains unmeasured (i.e., a vacant node), *S* is measured (i.e., a filled node) and affects the treatment A_2, the pretest Y_1, and the posttest Y_2. This *S* is a time-invariant covariate because the factor itself (for example, subjects' gender or race) does *not* change over time. At first glance, the time-invariant covariate *S* seems to not be a special variable but rather just a regular confounding variable. However, when it comes to the fixed effects approach, the key is in whether its impacts on the two measures Y_1 and Y_2 differ or not. Depending on the equality, we can classify *S* into either: i) a time-invariant covariate *whose impact is time-invariant* (if $\gamma_1 = \gamma_2$) or ii) a time-invariant covariate *whose impact is time-varying* (if $\gamma_1 \neq \gamma_2$). The former equality is a strong constraint, and the latter is more general. For example, although gender itself does not often change over time, the gender gap in income (i.e., the impact of gender on income) typically increases (or decreases) over time in many societies.

The necessity of including and controlling for time-invariant covariates in the fixed effects approach depends on the impacts of the covariates on the repeated outcome measures. If a time-invariant covariate belongs to the time-invariant covariate whose impact is time-*invariant*, then, as Allison (2009) has argued, it is not necessary to control for the covariate because the bias-removing mechanism by offsetting already takes it into account. In Figure 2.3B, we represent the gain score model (dashed arrows) with the time-invariant covariate *S* (but the offsetting principle also holds with the deviation score model). In this graph, the open non-causal paths between the treatment A_2 and the gain score *G*, together with their corresponding associations, are given by,

i) $A_2 \leftarrow U \rightarrow Y_1 \rightarrow G,\ \ -Var(U)\alpha_U\beta$;

ii) $A_2 \leftarrow U \rightarrow Y_2 \rightarrow G,\ \ Var(U)\alpha_U\beta$;

$$\text{iii)} \qquad A_2 \leftarrow S \rightarrow Y_1 \rightarrow G, \; -Var(S)\alpha_U\gamma_1;$$

$$\text{iv)} \qquad A_2 \leftarrow S \rightarrow Y_2 \rightarrow G, \; Var(S)\alpha_U\gamma_2.$$

In addition to path (i) and path (ii) offsetting each other, path (iii) and path (iv) also offset each other if the equality $\gamma_1 = \gamma_2$ holds, meaning that *S* is a time-invariant covariate whose impact is time-invariant.

However, if *S* is a time-invariant covariate whose impact is time-*varying*, which implies $\gamma_1 \neq \gamma_2$, the offsetting between path (iii) and path (iv) is not perfect and the resulting treatment effect estimate becomes biased, $Var(S)\alpha_U(\gamma_2 - \gamma_1) \neq 0$. Instead of relying on such imperfect offsetting, one may directly condition on *S* (because *S* is observed) to block the two non-causal paths via *S*, that is, path (iii) and path (iv). Then, the gain score model or the deviation score model allows for identification of the causal effect τ, despite $\gamma_1 \neq \gamma_2$. Contrary to the popular belief, this direct conditioning on *S* is straightforward. For example, after computing a gain score by $G = Y_2 - Y_1$, or a deviation score by $D = Y_2 - \left(\frac{Y_1 + Y_2}{2}\right)$, one simply regresses the gain score *G* (or *D*) on the treatment A_2 and the time-invariant covariate *S*.

Unfortunately, most of the literature on fixed effects models and the approach has not explicitly discussed whether a time-invariant covariate's impact on the repeated measures is time-invariant or time-varying (e.g., Allison, 2009; Clark & Linzer, 2015; Plümper & Troeger, 2007; Wooldridge, 2010; but see Halaby, 2004, as an exception). If the impact is time-invariant, the bias due to the omission of the covariate does not create any bias with the fixed effects approach. However, if the impact is time-varying, the bias cannot be removed by the offsetting mechanism alone, but it can be removed by directly blocking the non-causal paths. The graphical representations in Figure 2.3 help to make explicit the difference between a time-invariant covariate itself (i.e., a node) and its time-varying or time-invariant impacts on the repeated measures (i.e., arrows).

Controlling for Time-Invariant Covariates Using Dummy Variable Regression

However, it is true that, unlike the gain score or deviation score model, the dummy variable regression model cannot directly control for time-invariant covariates because individual dummies are perfectly *collinear* with time-invariant covariates (Allison, 2009; Wooldridge, 2010). Once the dummy variable ID_i is included in the pooled regression model, the time-invariant covariate cannot be included in the same pooled regression model regardless of its impacts on the repeated outcome measures. Probably, this is the source of the prevalent misleading belief that the fixed effects approach cannot include and control for any time-invariant covariates. However, dummy variable regression is just one of the several analytic models for identifying and estimating causal effects from fixed effects models, and other analytic models easily control for such covariates. Therefore, the inability to include and control for time-invariant covariates is not a limitation of the fixed effects approach itself.

Some researchers have tried to develop a way to resolve this inability in the dummy variable regression framework. Plümper and Troeger (2007, 2011) have proposed a method called the *fixed effects vector decomposition* (FEVD). This consists of three stages. Given the data depicted in Figure 2.3A, in the first stage, one runs a regular dummy variable regression that cannot include *S*: $\hat{Y}_{it} = bA_{it} + c_i ID_i + dT_t$. In the second stage, the obtained unit-specific effect c_i from the first stage is regressed on the time-invariant covariate *S*: $\hat{c}_i = a + bS_i$. Finally, in the third stage, the residualized unit-specific effect $\tilde{c}_i$ from the second stage, which has partialled out the impact of *S* (i.e., the residuals in the second regression model), plays the role of the individual dummy variables in the pooled regression model where the covariate *S* is included: $\hat{Y}_{it} = a + bA_{it} + c\tilde{c}_i + dT_t + eS_i$. Note that the lower-case *c* denotes the regression coefficient for the residualized unit-specific effects $\tilde{c}_i$. Also, the intercept *a* appears because the dummy variables ID_i are replaced with the univariate $\tilde{c}_i$. From this final regression model, one has the regression coefficient for *S* (i.e., the coefficient *e*), which cannot be obtained from the regular dummy variable regression model in the first stage.

The problem, however, as Greene (2011) has claimed and as Plümper and Troeger (2011) have agreed, is that the treatment effect estimate, which is the coefficient for A_{it} from the third stage, is indeed identical to the coefficient for A_{it} from the regular dummy variable regression model in the first stage. This means that the treatment effect estimate is still biased. In short, the FEVD is a way to somehow *include* time-invariant covariates in the dummy variable regression model, but it cannot really *control for* them to correct the bias in treatment effect estimates. The so-called *hybrid* method that uses group-mean centering for the treatment variable in the random effects framework (Allison, 2009) is the same in that it cannot correct the bias by including the time-invariant covariates. Although the obtained coefficients for the time-invariant covariates by the FEVD or the hybrid models might be useful in some contexts, they do not help to correct the bias in the treatment effect estimate. This is a severe limitation with respect to making a causal inference about the treatment effect of interest.

Indeed, there are simple ways to help researchers truly control for time-invariant covariates in the dummy variable regression framework. Results that are numerically equivalent to the gain score and the deviation score models that control for time-invariant covariates are obtained from a two-stage method that we shall refer to as the *dummy variable regression with the residualized treatment*. In the first stage, one regresses the treatment A_2 on the time-invariant covariate *S*: $\hat{A}_2 = a + bS_i$. In the second stage, the residual $\tilde{A}_2$ from the first regression, in which the impact of *S* is partialled out, is used to construct the treatment variable in a pooled dummy variable regression model: $\hat{Y}_{it} = b\tilde{A}_{it} + c_i ID_i + dT_t$, instead of the original treatment vector A_{it}. Then, the regression coefficient for the residualized treatment $\tilde{A}_{it}$ is numerically equivalent to the gain score and the deviation score estimators, in which the bias due to *S* is corrected.[4] Similarly, one may use matching. After matching the treatment and control cases on the values of the time-invariant covariates, the matched data sets can be used to run the regular dummy variable regression model.

Illustration

In order to demonstrate the effects of controlling for the time-invariant covariate *S*, we perform a simulation study. Based on the causal structure in Figure 2.3A, we generate data as follows:

[4] This is true for two-period data that we are considering in this paper. If the time points are more than two, the gain score estimator and the deviation score estimator are generally different (Wooldridge, 2010). Graphical representations also help in clarifying this difference, but in this paper, we do not extend and discuss this so that we can retain our focus on the basic two-period data. Note that even with more than two time points, the deviation score estimator is equivalent to the dummy variable regression estimator if *S* is not included in the deviation score model because these estimators are numerically identical without *S*.

$$
\begin{aligned}
&U \sim N(0,1);\\
&S \sim B(p=.5);\\
&A_2 = 1 \ \ if \ U + S + \varepsilon_A > 0; \ \ A_2 = 0 \ \ otherwise;\\
&Y_1 = U + \varepsilon_1;\\
&Y_2 = U + \tau A + S + \varepsilon_2,
\end{aligned} \tag{2.14}
$$

where ε_A, ε_1, and ε_2 are idiosyncratic error terms of each corresponding variable, following the standard normal distribution.[5] Note that we make the covariate *S* only affect Y_2; we do not make it affect Y_1. This highlights that the impact of *S* is obviously time-varying (i.e., $\gamma_1 = 0$ but $\gamma_2 = 1$). We generate 1,000 cases, and in each generated data set, we apply the analytic models and strategies previously described. The estimated treatment effects from each of those methods are saved and are averaged over 1,000 repetitions of data-generation. Finally, these expected treatment effects are compared with the true causal effect τ, which is set to .3 in the simulation.

The simulation results are summarized in Figure 2.4. The averages of the point estimates by different fixed effects methods are represented by hollow circles, and the corresponding plus or minus one standard deviations of the point estimates are represented by vertical lines through the circles. First, without controlling for the time-invariant covariate *S*, the three basic analytic models—the gain score model (“Gain”), the deviation score model (“Devi”), and the dummy variable regression model (“Dummy”)—produce biased effect estimates. However, if *S* is controlled for, through the gain score model (“Gain + S”) and the deviation score model (“Devi

[5] Strictly speaking, the data-generating model in Equation (2.14) is not a linear system and thus differs from the linear graph in Figure 2.3A. This is because in Equation (2.14) we consider that the treatment A_2 and the time-invariant covariate *S* are binary. In many causal inference studies, the treatment is binary (treatment group vs. control group) and many time-invariant covariates such as gender and race are binary or discrete rather than continuous. As this consideration does not affect our main arguments in the paper, we consider a more realistic simulation scenario that might differ, in a strict sense, from the linear graph in Figure 2.3A. Note that, however, Equation (2.14) reflects the conditional independence structures, implied by the graph in Figure 2.3A.

+ S"), the two estimators are unbiased. In contrast, due to the perfect collinearity, the dummy variable regression model cannot include *S* in the model, therefore, it is still biased ("Dummy + S"). Plümper and Troeger's (2007) fixed effects vector decomposition ("FEVD") and Allison's (2009) hybrid model ("Hybrid") also cannot correct the treatment effect estimates, although they somehow include *S* in the modified dummy variable regression models and provide the coefficients for *S*.[6] However, if the treatment is residualized ("Resid + Dummy") or cases are matched on *S* before running the dummy variable regression ("Match + Dummy"), the modified dummy variable regression models return unbiased effect estimates.[7]

Comparison with the Random Effects Approach

The comparison between the fixed effects and the random effects approaches has been widely discussed in the literature (Allison, 2009; Bell & Jones, 2015; Bollen & Brand, 2010; Clark & Linzer, 2015; Halaby, 2004; Lockwood & McCaffrey, 2007; Wooldridge, 2010, 2012). Although each method has its own advantages and disadvantages, it is not difficult to find in the literature strong exclusive preferences for either the fixed effects or the random effects approach. Allison (1994) claims, "the change-score estimator [i.e., the fixed effects approach] is nearly *always* preferable [to the random effects approach] for estimating the effects of events with nonexperimental data" (p. 181, emphasis added). In contrast, Gelman and Hill (2007) have

[6] Although both methods allow for the inclusion of *S*, the coefficients for *S* from the two methods are different. In the simulation, the coefficient for *S* by the FEVD method was .464 while the coefficient by the hybrid model was .198. The interpretation of these two coefficients are rather vague in terms of causal inference.

[7] As *S* is binary in the simulation, we use exact matching for "Match + Dummy." We split the total sample into two, based on the values of *S*, and run the separate dummy variable regression models using each sample. The final effect estimate is obtained by the weighted average of the two separate treatment effects (weighted by the frequency of *S*).

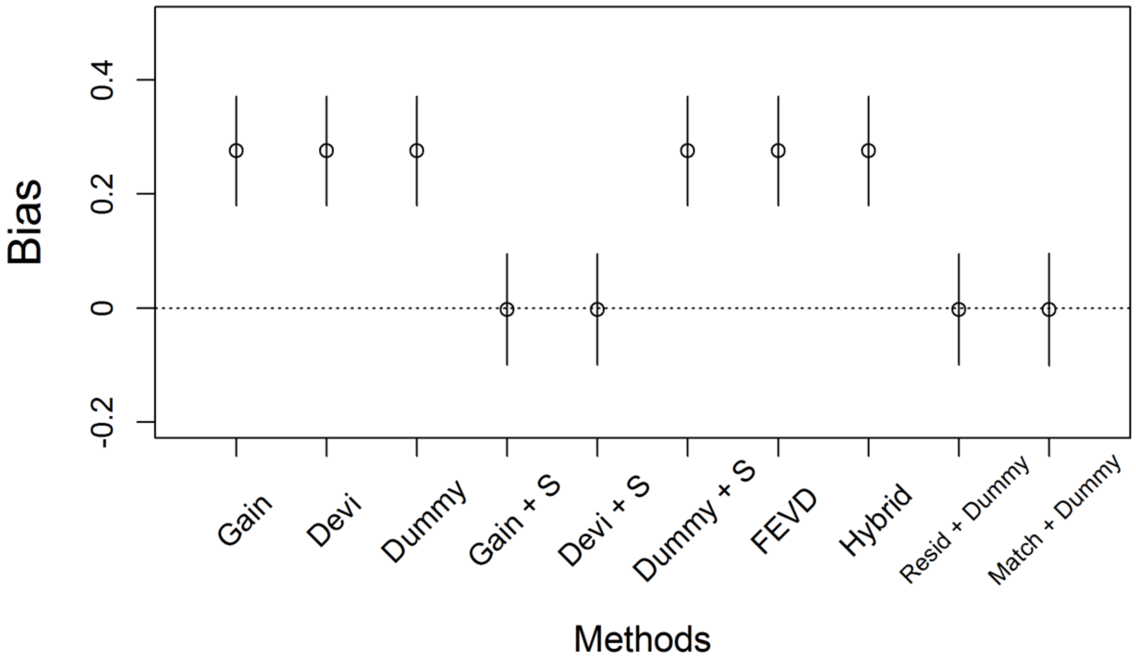

FIGURE 2.4. *Remaining bias by different analytic models for a fixed effects model where S is present. Gain = Simple gain score; Devi = Simple deviation score; Dummy = Dummy variable regression; Gain + S = Gain score controlling for S; Devi + S = Deviation score controlling for S; Dummy + S = Dummy variable regression controlling for S; FEVD = Fixed effects vector decomposition (Plümper & Troeger, 2007); Hybrid = Hybrid model (Allison, 2009); Resid + Dummy = Dummy variable regression with the residualized treatment, partialling out S; Match + Dummy = Dummy variable regression with the matched sample on S.*

written, "[o]ur advice (elaborated upon in the rest of this book) is *always* use multilevel modeling ('random effects')" (p. 246, emphasis and parentheses in original).

Here, we focus on the bias-removing mechanism using the basic random effects approach, specifically, the random intercepts models. Although it is well documented algebraically that using the random effects approach results in biased effect estimates if the unit-fixed effects are *correlated* with the treatment (Lockwood & McCaffrey, 2007; Woodridge, 2010), the literature has not explicitly investigated the corresponding graphical representations and the causal interpretation of the random effects approach. As we will show, graphical representations help us in understanding how the random effects approach removes confounding bias and why it is inherently limited to being used for causal identification in observational studies.

Imperfect Bias-Offsetting by the Random Effects Approach

In Figure 2.5A, we present the graphical representation of the bias-removing mechanism of the random effects approach. Given the basic two-period data under the fixed effects model in Figure 2.1B, the analytic structure of the random effects approach is represented by dashed arrows. The added structure is indeed identical to the structure of the deviation score model (see Figure 2.2B), but the path coefficients on the two arrows, $Y_1 \to w$ and $Y_2 \to w$, are different such that, instead of $+{}^{1}\!/_{2}$, they are $+{}^{\lambda}\!/_{2}$. The parameter λ is what has become known in the multilevel literature as the *reliability*, which produces the so-called *shrinkage estimator* (Raudenbush & Bryk, 2002). Multilevel researchers conceive that the observed average of the repeated measures Y_1 and Y_2 for each *i* consist of two parts: *i*'s true unknown score and random error. Thus, the parameter λ quantifies "the ratio of the true score […] variance, relative to the observed score or total variance of the sample mean" (Raudenbush & Bryk, 2002, p. 46). Note that in the graph, we refer to the average node as *w* instead of *m*, and we refer to the resulting deviation score node, computed by $Y_2 - w$, as *qD* (denoting the *quasi-deviation score*) instead of *D*. Wooldridge (2010) clarified this computation procedure of random effects estimators (see also Lockwood & McCaffrey, 2007), and referred to it as *quasi-time demeaning*. Remember that the time demeaning model in the econometrics literature is what we here refer to as the deviation score model.

In Figure 2.5A, the open non-causal paths between the treatment A_2 and the quasi-deviation score *qD*, together with the corresponding associations, are given by

i) $A_2 \leftarrow U \rightarrow Y_2 \rightarrow qD,\ Var(U)\alpha\beta$;

ii) $A_2 \leftarrow U \rightarrow Y_2 \rightarrow w \rightarrow qD,\ -Var(U)\dfrac{\alpha\beta}{2}\lambda$;

iii) $A_2 \leftarrow U \rightarrow Y_2 \rightarrow w \rightarrow qD,\ -Var(U)\dfrac{\alpha\beta}{2}\lambda$.

However, note that the sum of all the three non-causal associations, $Var(U)\alpha\beta(1-\lambda)$, does *not* equal zero unless the reliability becomes one ($\lambda = 1$). This is why the random effects approach, unlike the fixed effects approach, have a fundamental limitation in eliminating confounding bias by offsetting. As the random effects approach exploits the shrinkage estimator, which is a weighted average of the observed sample means (by *i*) and the grand mean (across *i*) by the reliability λ, the bias-offsetting mechanism is inherently imperfect even when the impacts of *U* on Y_1 and Y_2 are identical.

Comparison of Two Approaches in Experimental and Non-Experimental Studies

Despite the imperfect bias-offsetting by the random effects approach, Bell and Jones (2015) have claimed that “[i]f the assumptions made by RE [i.e., random effects] models are correct, RE would be the preferred choice because of its greater flexibility and generalizability, and its ability to model context […]” (p. 134). Graphical representations also help in clarifying

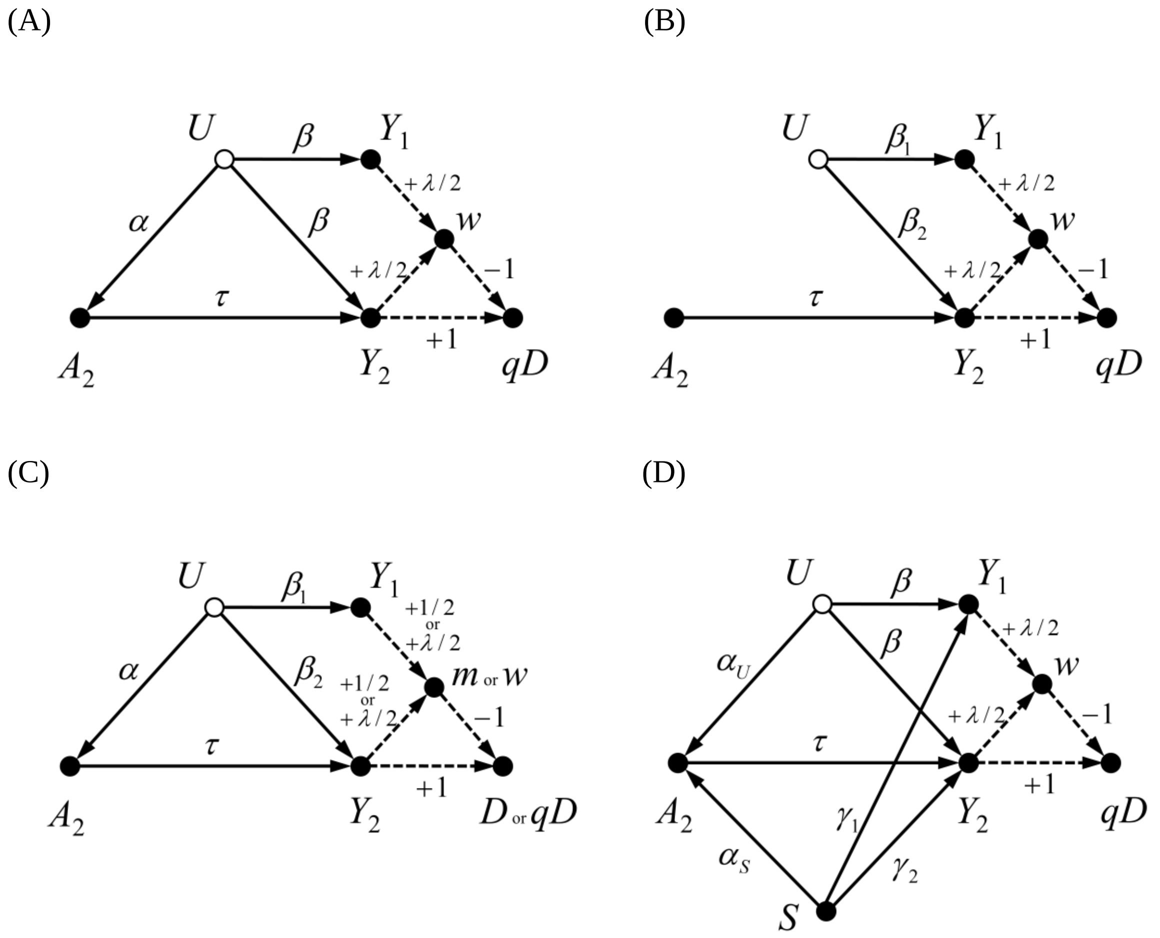

FIGURE 2.5. *Graphical representations, (A) adding the random effects model structure (dashed arrows) to the data in Figure 2.1B, (B) adding the random effects model structure to RCT data, (C) adding the fixed and random effects model structures to observational data, and (D) adding the random effects model structure to the data in Figure 2.3A, where time-invariant covariate S is present.*

the conditions under which the random effects approach does not make any bias and thus may be preferable over the fixed effects approach due to other benefits mentioned by Bell and Jones. As previously considered, in the basic fixed effects model in Equation (2.1), $Y_{it} = \tau A_{it} + \theta_i + \varepsilon_{it}$, we assume that the fixed effect is correlated with the treatment, $Cov(A_{it},\theta_i)\neq 0$. This assumption is encoded in the graphs as the arrow $U\rightarrow\ A_2$ because $\theta_i=\beta U_i$. If this covariance is zero, implying that the treatment is independent of the fixed effect, then the random effects approach does not result in any bias (Allison, 1994; Lockwood & McCaffrey, 2007; Halaby, 2004; Wooldridge, 2010). This assumption $Cov(A_{it},\theta_i)=0$ is occasionally referred to as the *random effects assumption* (Bell & Jones, 2015).

In Figure 2.5B, we describe the random effects assumption by deleting the arrow $U\rightarrow$ A_2. If this is the data-generating model, then the random effects estimators are unbiased. This is simply because there are no open non-causal paths between A_2 and *qD*. The only open paths are the following two causal paths

i) $A_2\rightarrow Y_2\rightarrow qD,\ Var(A_2)\tau$;

ii) $A_2\rightarrow Y_2\rightarrow w\rightarrow qD,\ -\dfrac{Var(A_2)\tau}{2}\lambda$.

The sum of the two paths is $Var(A_2)\tau\left(\dfrac{2-\lambda}{2}\right)$ and, as for the deviation score estimators, in order to recover the original causal quantity, one should multiply this quantity by ${}^{2}\!/\!_{(2-\lambda)}$ in a two-period panel data. This graph-based argument verifies that under the random effects assumption, $Cov(A_{it},\theta_i)=0$, the random effects approach can result in unbiased effect estimates.

In fact, the graph in Figure 2.5B, especially the structure with the solid arrows, is the data-generating model for randomized controlled trials (RCTs), where the treatment variable is not confounded by other unmeasured variables (Steiner, Kim, Hall, & Su, 2017). This assumption, or equivalently the independence assumption conditional on other measured covariates (i.e., unconfoundedness; Imbens & Rubin, 2015), is one of the strongest assumptions for making a causal inference using observational data. If this assumption holds, then the random effects approach as well the fixed effects approach and even a naïve approach (e.g., simply regressing the outcome on the treatment without any other conditioning variables), all result in unbiased effect estimates because the treatment-outcome relationship is unconfounded in RCT data. This can be verified in the graph by the absence of the open non-causal paths between the treatment and the outcome variables. Particularly, with RCT data, the key constraint for being a fixed effects model—the equal impacts of *U* on Y_1 and Y_2, (i.e., the common trend or fixed effects assumption)—is not even required in order to apply the fixed effects approach. Notice that in Figure 2.5B, we use the different parameters β_1 and β_2 for the arrows $U \rightarrow Y_1$ and $U \rightarrow Y_2$, respectively, indicating $\beta_1 \neq \beta_2$. Even with this data-generating model where the key assumption is not met, the fixed effects estimators would be unbiased because there are no non-causal paths between the treatment A_2 and the deviation score *D*.

Now consider a general observational study where the treatment is confounded by *U* and the impacts of *U* on Y_1 and Y_2 are different, $\beta_1 \neq \beta_2$. This is plausible case in many observational studies in practice. The data-generating model is presented in Figure 2.5C by solid arrows. With this kind of data, one might wonder which analytic approach (i.e., random effects vs. fixed effects) would be preferred. In Figure 2.5C, the two analytic options are displayed together in the graphical representation. Depending on which is selected, different path coefficients are specified, and the confounding bias will be differently offset. Note that, due to $\beta_1 \neq \beta_2$, the fixed effects approach (more exactly, the deviation score model) cannot eliminate all the bias. The random effects approach also retains some amount of bias because it inherently uses the imperfect offsetting mechanism. In this case, it is never clear which approach would result in less bias and thus which one should be generally preferred. With this observational data in Figure 2.3C, even the Hausman test (Hausman, 1978) does not help us choose one approach. Although the fixed effects estimator and the random effects estimator would differ, this does not indicate that the fixed effects approach should be preferred (Clark & Linzer, 2015). This is because the key assumption $\beta_1 = \beta_2$ for being a fixed effect model does not hold with this type of general observational data. In this case, depending on the specific values for the two parameters, it is possible that random effects estimators are much less biased than fixed effects estimators.

Illustration

One might wonder whether our graphical representations for the random effects approach in Figure 2.5 correctly represent what happens in the random effects estimation. We refer interested readers to Wooldridge (2010) or Lockwood and McCaffrey (2007) for the algebraic derivation. Here, we provide numerical evidence to support this derivation. Using the same data-generating model in Equation (2.14), we apply two random intercepts models to the simulated data. Two random effects models are different depending on whether or not the time-invariant covariate *S* is controlled for:

$$\begin{aligned} Level1: &\quad Y_{it} = \pi_{0i} + \pi_{1i} T_{it} + e_{it}, \\ Level2: &\quad \pi_{0i} = \delta_{00} + r_{0i}, \\ &\quad \pi_{1i} = \delta_{10} + \delta_{11} A_{2i}, \end{aligned} \tag{2.15}$$

and

TABLE 2.1.

Comparison between random effects estimators and quasi-deviation score estimators

	Random effects estimators	Reliability λ	Quasi-deviation score estimators
No controlling for *S*	1.2270515	0.5667732	1.2270515
Controlling for *S*	0.9437824	0.6145112	0.9437824

$$\begin{aligned} &Level1: \quad Y_{it} = \pi_{0i} + \pi_{1i}T_{it} + e_{it}, \\ &Level2: \quad \pi_{0i} = \delta_{00} + r_{0i}, \\ &\qquad\qquad\; \pi_{1i} = \delta_{10} + \delta_{11}A_{2i} + \delta_{12}S_i, \end{aligned} \tag{2.16}$$

where *S* is controlled for in Equation (2.16). The two models first specify the individual time trends by the level-1 models and then investigate the treatment group difference in the slopes (π_1) by the level-2 models. The coefficient δ_{11} is the random effects estimator of interest. To specify the models, we use the function *lmer* in the R-package *lme4* (Bates, Maechler, Bolker, & Walker, 2015). In each analysis, we extract the reliability λ and use it to compute the quasi-deviation score and apply two quasi-deviation score models, again depending on whether or not the time-invariant covariate *S* is controlled for. The graphical representation for the quasi-deviation model controlling for *S* is described in Figure 2.5D. Once the quasi-deviation score is computed, we regress it on the treatment A_2 and the time-invariant covariate *S*.

The results are summarized in Table 2.1. From the simulation (1,000 times of data-generating processes), the random effects estimator when *S* is not controlled for, as in Equation (2.15), is 1.23. Using the obtained reliability λ in each analysis, we compute the quasi-deviation score estimator and compare it with the random effects estimator. The two estimators are numerically identical and we present the results up to seven decimal places as presented in the row labeled “No controlling for *S*” in Table 2.1. This numerical equivalence between the random effects estimator and the quasi-deviation score estimator still holds when *S* is controlled for (the "Controlling for *S*" row). The results support the validity of our graphical representations, interpreting the random effects estimation as using the quasi-deviation score model.

DISCUSSION

In this paper, we develop graphical representations for the fixed effects approach that can be useful for psychological studies. The key causal identification process of the approach is clearly explained through linear structure models and corresponding graphical models. Unlike the usual methods such as regular regression or matching that aim to block the non-causal path(s) between the treatment and the outcome, the fixed effects approach (including the gain score model, the deviation score model, and the dummy variable regression model) creates another non-causal path(s) and tries to offset the original non-causal path(s). This unique bias-removing mechanism requires its own identification assumption (i.e., the common trend or fixed effects assumption) that is different from the assumption for the regular regression or matching (i.e., unconfoundedness). Importantly, we show that, contrary to the widespread belief, the inability to control for time-invariant covariates is definitely not a limitation of the fixed effects approach. It may be a limitation of the standard dummy variable regression model, but we also provide simple remedies so that the covariates can be controlled by using the residualized treatment or matching.

Revealing the underlying causal structures helps in resolving the prevalent misunderstanding and confusion about the fixed effects and other related approaches. At first glance, the gain score model and the fixed effects approach seem to be two separate methods. But, once they are represented by graphs, the strong resemblance and relationship become apparent. The gain score model, which is popular but nonetheless has long been criticized within the psychological literature (e.g., Campbell & Erlebacher, 1970; Cronbach & Ferby, 1970), can be understood as one of the several analytic models that identifies causal effects from the data generated by fixed effects models. Note that the gain score model is indeed a difference-in-differences (DiD) model. It is interesting that the problem of time-invariant covariates has been a long-standing puzzle in the literature about fixed effects approach (Beck, 2011; Bell & Jones, 2015; Breusch, Ward, Nguyen, & Kompas, 2011a, 2011b; Greene, 2011; Plümper & Troeger, 2007, 2011) while the same problem has not occurred in the literature about DiD. Rather, using matching to control for pre-intervention covariates before running a DiD model is a common approach (e.g., Stuart et al., 2014; Wichman & Ferraro, 2017). If the DiD part (i.e., the gain score model) is replaced with the dummy variable regression model, the resulting approach becomes our proposed dummy variable regression using the matched data that allows for controlling for time-invariant covariates in the dummy variable regression framework.

Probably, the most important implication of this paper for researchers and analysts studying causal inferences is the importance of having strong substantive knowledge about the data-generating process. This structure was represented by solid arrows in our graphs. Given this knowledge, researchers can carefully consider possible analytic models—model structures which were represented by dashed arrows in our graphs. An analytic model will eliminate bias within a given data set, but with other sets of data, it may completely fail to eliminate bias and even become severely misleading. There is no generally superior analytic method or model regardless of how the data were generated. For example, the popular random effects approach cannot eliminate all confounding bias if the treatment is associated with the fixed effects.[8] This claim is not new (see Allison, 1994, 2009; Wooldridge, 2010, 2012), but our graphs make this point explicit. From our graphs (e.g., Figures 2.2B and 2.5A), one can easily see why random effects estimators are biased (imperfect offsetting) while fixed effects estimators are unbiased (perfect offsetting). Although we do not discuss this in the current paper, it is possible that the gain score model eliminates all the bias but the deviation score model leaves bias (and vice versa). Strong knowledge about the data-generating process will guide researchers to choose the best analytic method or model. Graphical models will significantly facilitate this decision-making process.

[8] All the random effect estimators in Table 2.1 are much greater than the true causal effect, which was set to .3 in the simulation.

STUDY 3

CAUSAL IDENTIFICATION USING DIFFERENCE-IN-DIFFERENCES IN MEDIATION ANALYSIS

Abstract

Causal identification of direct and indirect effects requires a stringent, no unmeasured mediator-outcome confounding assumption even when the treatment is randomized. We develop a difference-in-differences in mediation analysis (DiDiM) approach by extending the standard difference-in-differences to the mediation context. This extension allows causal identification of direct and indirect effects in the presence of mediator-outcome confounding. The unique bias-removing mechanism of the proposed approach is highlighted by showing that this approach may even identify direct effects when it is impossible to identify either total or indirect effects because the treatment is not randomized. Although this approach requires its own strong identification assumption, we also propose a method to adjust the DiDiM estimators for the bias due to the violation of the key assumption. We apply these methods to real data sets and show that the DiDiM and the adjusted DiDiM methods substantially mitigate the mediator-outcome confounding bias and outperform the standard mediation methods.

INTRODUCTION

As in many scientific fields, identifying causal mechanisms is crucial for educational research. This is so not only because of theory development but also because of practical interest; for example, for policy evaluation. Without understanding the process of an educational intervention, it is difficult to properly evaluate a given policy (Caro, 2015; Kaplan, 2009). For instance, although an education accountability act, such as the No Child Left Behind (NCLB) or the Race to the Top, may increase student achievement in some subjects (Dee & Jacob, 2011; Wong, Cook, & Steiner, 2009), it is unclear what mechanisms lead to such results. These results may be due to teachers' and school administers' positive efforts, as (implicitly) intended by policymakers, or due to educators' undesirable reactions under pressure such as teaching to the test, excluding low-achieving students from testing, and even test-score manipulation (Nichols & Berliner, 2005). Obviously, if much of the policy's effect is made by such undesirable processes or mechanisms, the policy should be re-evaluated despite the overall positive effect.

Causal mediation analysis aims to uncover such underlying causal mechanisms between treatment (e.g., NCLB) and outcome (e.g., achievement score). The standard approach to causal mediation defines causal mechanisms as *direct* and *indirect* effects (Imai, Keele, & Yamamoto, 2010b). Identifying causal mechanisms, then, involves discerning how to decompose the total effect of the treatment into the direct effect and the indirect effect, depending on whether the effect is mediated by an intermediate variable between treatment and outcome, called a *mediator* (e.g., test-score manipulation). Since Baron and Kenny's (1986) influential work (also Judd & Kenny, 1981), many statistical approaches for decomposing total effects into direct and indirect effects have been developed, elaborated, and extended (e.g., Glynn, 2012; Hayes, 2009;

MacKinnon, 2008; MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002; Preacher & Hayes, 2008; Preacher, Rucker, & Hayes, 2007).

Importantly, with the recent development of causal frameworks such as potential outcomes and graphical causal models, causal interpretation and identification of mediation effects have been substantially improved (e.g., Imai, Keele, & Tingley, 2010a; Imai et al., 2010b; Pearl, 2014; VanderWeele, 2015). Currently, conditions under which causal direct and indirect effects can be identified from observed data are well documented in the literature (Imai et al., 2010a, 2010b; Pearl, 2014; Ten Have & Joffe, 2010; VanderWeele, 2010, 2015). In order to identify such effects, there is no unmeasured confounding of: i) the treatment-outcome, ii) the treatment-mediator, and iii) the mediator-outcome relationships.[1] Among these, the third assumption in particular has remained "the fundamental difficulty in the causal mediation analysis" (Imai et al., 2010a, p. 310). This is because this assumption (i.e., no unmeasured mediator-outcome confounding) is still in doubt even in randomized controlled trials (RCTs), which are often considered the *gold standard* of causal inference (the other two confounding assumptions would hold in RCTs).

In this paper, we develop a new approach to causal mediation analysis that allows for identification of mediation effects in the presence of unmeasured mediator-outcome confounding. We do this by extending the standard difference-in-differences method to the mediation context. We clarify the conditions under which the proposed difference-in-differences in mediation analysis (DiDiM) approach identifies causal mediation effects. Interestingly, even

[1] Another assumption is that there is no unmeasured or even *measured* variable that is affected by the treatment and that confounds the mediator and the outcome (VanderWeele, 2010, 2015). In this paper, we consider that this assumption is met and focus on the unmeasured mediator-outcome confounding assumption.

when the treatment is not randomized, it is possible that the DiDiM identifies causal direct effects but not total nor indirect effects. After formalizing the key identifying assumption, we also develop a method to adjust DiDiM estimators for the violation of this assumption. These methods will be applied to real data sets, and we show evidence that the DiDiM and the adjusted DiDiM methods substantially mitigate the unmeasured mediator-outcome confounding bias. Before proceeding to our solution, we will first describe in more detail the problem of unmeasured mediator-outcome confounding.

THE PROBLEM OF MEDIATOR-OUTCOME CONFOUNDING

Understanding Unmeasured Mediator-Outcome Confounding

Unmeasured mediator-outcome confounding is one of the major challenges in causal mediation analysis (Imai et al., 2010a; Small, 2012; VanderWeele, 2010). This is illustrated in Figure 3.1A. The graph describes a causal system where the treatment *A* causally affects the outcome *Y*, and part of the effect is mediated by the mediator *M*. We consider that the treatment is binary and randomized, and for ease of exposition, the mediator and outcome variables are continuous.[2] The randomization of *A* is encoded in the graph as the absence of incoming arrows into *A*; that is, there are no confounding variables that affect the treatment *A*. Graphically, intervening on a variable (e.g., randomizing) cuts off all incoming arrows into the intervened variable (Pearl, 1993, 2009). However, randomizing the treatment alone does not eliminate any potential confounding between the mediator and the outcome. In Figure 3.1A, the variable *U*,

[2] Viewing the mediator and outcome as continuous will significantly simplify the algebra. But, the basic principles we present in the paper can be extended to cases of discrete mediators and outcomes.

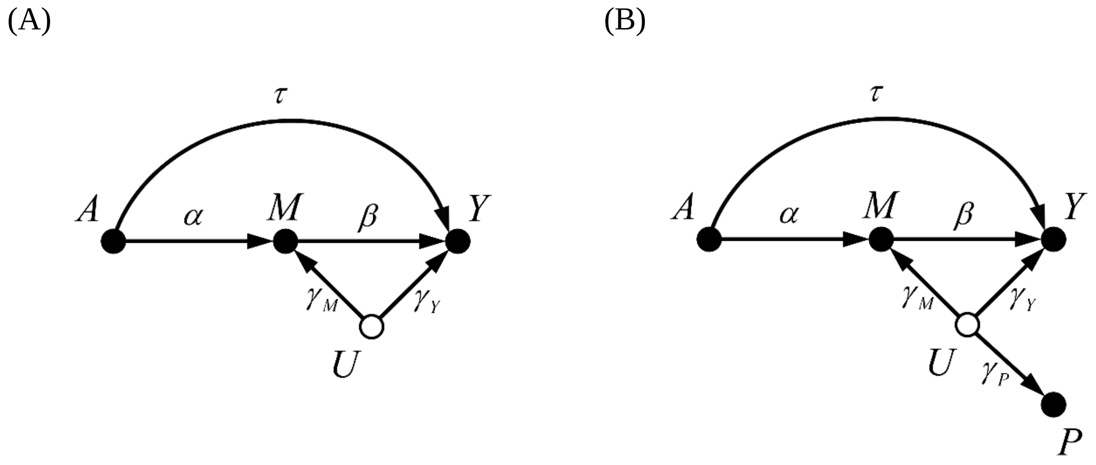

FIGURE 3.1. *Basic causal mediation structures in randomized experimental studies. (A) Unmeasured mediator-outcome confounding variable U is present. (B) Pretest measure P is added to graph (A).*

outcome confounding, but, it remains *blocked* at the collider *M* (unless conditioning on *M*; see Elwert & Winship, 2014; Pearl, Glymour, & Jewell, 2016). Therefore, this third non-causal path does not hinder us from identifying the total effect of *A* on *Y*, which is transmitted via two causal direct and indirect paths. For more explanations of (un)blocking paths and the formal principle called *d-separation*, see Pearl (2009) or Pearl et al. (2016).

However, the mediator-outcome confounding *does* hinder causal identification of mediation effects. To decompose the total effect into the direct and indirect effects, one needs to condition on the mediator *M* so as to isolate the direct path (i) from the indirect path (ii). Conditioning on *M* blocks indirect causal path (ii), and this is what researchers intend to do so in mediation analysis. However, this also generates an unfortunate consequence. The third non-causal path is then open because the collider *M* is conditioned on. This means path (iii) starts transmitting a spurious association between *A* and *Y*. As a result, what researchers observe after they have conditioned on *M* is not only due to the direct causal path (i) but also due to the non-causal path (iii). This means the direct effect of *A* on *Y*, transmitted only via the direct path (i), is not generally identified from data by simply conditioning on the mediator.

This graphical intuition about mediator-outcome confounding is verified algebraically. For ease of exposition and without loss of generality, in this paper we assume that all unmeasured variables (vacant nodes), including omitted random disturbances, have unit-variances, thus, in Figure 3.1A, $Var(U) = 1$. Each Greek letter lying on the arrows represents the constant path coefficients that do not change across units, implying that we consider linear systems. In the graph, the effect of the treatment *A* on the outcome *Y* is direct ($A \rightarrow Y$: τ) and indirect via the mediator *M* ($A \rightarrow M \rightarrow Y$: $\alpha \times \beta$), and the total effect is simply the sum of the direct and indirect effects, $\tau + \alpha\beta$. To identify the total effect, one regresses *Y* on *A* (or compares the group means in the outcome using a *t*-test). The population regression coefficient for *A* is then

$$b_{YA} = \frac{Cov(Y, A)}{Var(A)} = \tau + \alpha\beta , \tag{3.1}$$

by applying Wright's (1921) path-tracing rules (see Pearl, 2013, for details) or simple algebra.[3] So, the regression coefficient exactly corresponds to the true total effect. However, if one controls for *M* such as by regressing *Y* on *A* and *M*, the population partial regression coefficient for *A* is given by

$$b_{YA|M} = \tau - \frac{\alpha\gamma_M\gamma_Y}{Var(M) - Var(A)\alpha^2} , \tag{3.2}$$

[3] See Appendix D for derivations of all regression estimator formulae hereafter.

which differs from the true direct effect τ. The second term of Equation (3.2) represents the spurious association transmitted via the third non-causal path, $A \rightarrow M \leftarrow U \rightarrow Y$, which is opened due to conditioning on M. This correspondence can be seen from the fact that the numerator of the bias term consists of the product of α, γ_M, and γ_Y; each of which is the path coefficient along path (iii).[4] The denominator $Var(M) - Var(A)\alpha^2$, or a scaling factor, which occurs due to partialling out the variations among regressors, is not directly depicted in graphs. Equation (3.2) verifies that what we observe after conditioning on M is a mixture of the associations via path (i) and path (iii). If U is not a confounding variable such that $\gamma_M \gamma_Y = 0$, then Equation (3.2) reduces to $b_{YA|M} = \tau$, implying that one identifies the direct effect simply by conditioning on M. However, if there is any unmeasured confounding variable between M and Y, the regression coefficient, aimed at identifying the direct effect of A on Y, is always biased.

Dealing with the Mediator-Outcome Confounding

Standard mediation approaches collect covariates in order to capture the unmeasured mediator-outcome confounding and try to deconfound the mediator-outcome relationship by conditioning on the covariates (see sequential ignorability in Imai et al., 2010b). "[I]f we are interested in mediation analysis, then we *must control for* mediator-outcome confounders" (VanderWeele, 2016, p. 21, emphasis added). This covariate adjustment, however, is problematic, as in causal identification of total effects, because we cannot guarantee that

[4] It is interesting to note the role of α in the formula. In Equation (3.2), as the absolute value of α increases, the bias term is strongly amplified. This is similar to what Pearl (2010) referred to as *bias amplification* (also see Steiner & Kim, 2016). Equation (3.2) shows that the bias amplification also occurs in the mediation context.

measured covariates will cover all the unmeasured confounding variables. In practice, there is almost always remaining unmeasured confounding between the mediator and the outcome. Moreover, it is possible that controlling for covariates even increases the bias in total effect estimates (Myers et al., 2011; Pearl, 2010, 2011; Steiner & Kim, 2016), and this may also occur in mediation analysis.

Let us focus on a single covariate, which also will be used later in a different way for our new approach. Probably, the best single covariate for covariate adjustment is the pretest measure of the outcome, simply referred to as *pretest* (Cook & Steiner, 2010; Shadish, Cook, & Campbell, 2002). In general, such a pretest is considered as a *proxy* of the unmeasured confounding variable, and we depict the pretest *P* as a variable that is affected by *U* in Figure 3.1B. However, in general, this pretest cannot eliminate all the mediator-outcome confounding bias due to *U*. Graphically, this is straightforward because the pretest *P* does not sit on the confounding or backdoor path between *M* and *Y*. Conditioning on *P* cannot fully block the path and establish the necessary conditional independence for deconfounding the relationship between *M* and *Y*.

This also can be verified using algebra. Regressing *Y* on *A*, *M*, and, additionally, *P* (a method that also controls for the proxy of unmeasured confounding), the population partial regression coefficient for *A* is given by

$$b_{YA|MP} = \tau - \frac{\alpha\gamma_M\gamma_Y(1-\rho)}{Var(M) - Var(A)\alpha^2 - \gamma_M^2\rho}, \tag{3.3}$$

where $\rho = \frac{\gamma_P^2}{1+\gamma_P^2}$. This ρ can be interpreted as the pretest's reliability in terms of the unmeasured confounding variable *U* shared with *Y*. Note that the omitted random disturbance of

P has unit-variance, so the total variance of *P* is $Var(P)=1+\gamma_P^2$. If the reliability is exactly one, implying that *P* is a "perfect" proxy of *U*, then the bias term in Equation (3.3) goes away, and the regression coefficient becomes the unbiased estimate of the direct effect τ . However, in general, with any realistic ranges of the reliability, the regression coefficient is always, to some extent, biased. The consequence of controlling for many covariates can be also understood with the concept of reliability. If a set of covariates helps to increase the overall reliability (in terms of *U*), it may reduce some of the mediator-outcome confounding bias. Nonetheless, it likely contains some amount of bias because the reliability will never be perfect in real-world situations even if hundreds of covariates are collected. Also, note that in Equation (3.3), the reliability ρ plays a role in amplifying the bias term by reducing the magnitude of the denominator (i.e., the subtractive term $-\gamma_M^2\rho$). This suggests that the increase of the reliability, which can occur through the inclusion of as many covariates as possible, may not always help to decrease the overall unmeasured mediator-outcome confounding bias.

Although this covariate adjustment is the most popular method in practice, and will be later compared with our DiDiM approach, it is worth discussing two other recently proposed approaches to circumvent the problem of unmeasured mediator-outcome confounding.[5] This will highlight the difference between the DiDiM and other approaches and will clarify the conditions under which each method would be applicable in practice.

[5] Another approach we do not discuss here is sensitivity analysis, probing the range of possible mediation effects along with the degree of the violation of the unmeasured mediator-outcome confounding (e.g., Ding & VanderWeele, 2016; Imai et al, 2010a, 2010b; VanderWeele, 2010). Strictly speaking, this useful method does not help us directly identify causal mediation effects in the presence of unmeasured mediator-outcome confounding.

First, Dunn and Bentall (2007), Ten Have et al., (2007), and Small (2012) have proposed and elaborated an *instrumental variables* approach. The original purpose of using instrumental variables was to avoid unmeasured treatment-outcome confounding in cross-sectional or non-mediation studies, and the similar rationale would also hold for unmeasured mediator-outcome confounding in mediation studies. They consider a baseline covariate that interacts with the randomized treatment on the mediator. This interaction is then used as a conditional instrumental variable, conditional on the treatment and the covariate, with respect to the mediator-outcome relationship. Using two-stage least squares, one first regresses the mediator *M* on the treatment *A*, the baseline covariate *X*, and the interaction between them (i.e., cross-product term *AX*). In the second stage, one regresses the outcome *Y* on the treatment *A*, the covariate *X*, and the predicted mediator $\hat{M}$ from the first stage regression. Dunn and Bentall (2007) and Ten Have et al. (2007) showed that this strategy identifies unbiased mediation effects even in the presence of unmeasured mediator-outcome confounding. They assumed that the effect of the mediator on the outcome and the direct treatment effect on the outcome are constant across units, but Small (2012) slightly relaxed the restriction by making a weaker conditional independence assumption.

The problem with this instrumental variables approach is the same as the problem with the standard instrumental variables approach; it is challenging to find a proper instrumental variable, that is, an interacted covariate, that satisfies the required assumptions. First, the presence of the interaction between the treatment *A* and the baseline covariate *X* on the mediator *M* is the *first-stage* assumption, which requires that instrumental variables be strongly associated with the treatment (in our context, the mediator *M*). If the association, that is, the impact of the interaction, is weak, the resulting so-called *weak*-instruments can lead to large variance estimates. Second, the instrumental variables approach also requires the absence of the pairwise interactions between the treatment, the mediator, and the baseline covariate on the outcome. This is related to the *exclusion restriction* assumption. If this strong absence of interactions is not met, the interaction term *AX* can be associated with the outcome *Y* not via the mediator *M* (conditional on *A* and *X*), therefore, the resulting mediation estimators are biased.

Second, He, Wu, Zhang, and Geng (2016) have recently proposed an approach relying on the different nonlinearity degree of the mediator and outcome equations. Their approach requires that the degree of equation nonlinearity for the treatment-mediator be higher than that for the treatment-outcome. For example, the functional form of the mediator *M* follows a quadratic model with respect to *A*, such as $M = \theta_1 A + \theta_2 A^2 + U + e_M$, while the functional form of the outcome *Y* follows a linear model with respect to *A* and *M*, such as $Y = \pi_1 M + \pi_2 A + U + e_Y$, where *U* is the unmeasured mediator-outcome confounding variable and e_M and e_Y are the random disturbances of each.

As the conditional expectation of *Y* on *A* is expressed with $E[Y \mid a] = \pi_1 E[M \mid a] + \pi_2 a$, by comparing the conditional expectations at different treatment levels, one derives the following equation:

$$\begin{bmatrix} E[M \mid a_1] - E[M \mid a_2] & a_1 - a_2 \\ \vdots & \vdots \\ E[M \mid a_{k-1}] - E[M \mid a_k] & a_{k-1} - a_k \end{bmatrix} \begin{bmatrix} \pi_1 \\ \pi_2 \end{bmatrix} = \begin{bmatrix} E[Y \mid a_1] - E[Y \mid a_2] \\ \vdots \\ E[Y \mid a_{k-1}] - E[Y \mid a_k] \end{bmatrix}, \tag{3.4}$$

where $a_1, \cdots, a_k$ are *k* distinct values of the treatment *A*. If the $(k-1) \times 2$ matrix in the left-hand side has full column rank and thus is invertible, the parameters π_1 and π_2 are identifiable because two conditional expectation vectors are observed. He et al. (2016) extend this approach to more general cases of nonlinear outcome models and also propose an idea that may relax the nonlinearity assumption using instrumental variables.

The use of this approach, however, may be limited in practice because in many causal studies, researchers have two values in the treatment variable, indicating treatment and control group memberships. In this case, the $(k-1)\times 2$ matrix in Equation (3.4) is not invertible so that we cannot identify the parameters π_1 and π_2 (He et al., 2016). At least three distinct values are required for the identification. Depending on what the given research questions are, this may hold, but this type of requirement is definitely not a good fit for many causal inference questions in the literature.

In sum, even in RCTs, unmeasured mediator-outcome confounding prevents us from identifying causal mediation effects. By using other variables such as pretest measures of the outcome or instrumental variables for the mediator, one may try to circumvent the problem. However, the unreliability of the pretest is unavoidable and good instrumental variables are rare in practice. Also, researchers often compare two treatment groups, so some mathematical techniques for solving equations might not be applicable. In the next section, we present an alternative to these mediation approaches by using difference-in-differences.

Causal Mediation Analysis Using Difference-in-Differences

"Difference-in-Differences in Mediation Analysis" in Experimental Studies

Graphically speaking, the difficulty of unmeasured mediator-outcome confounding is that by conditioning on the mediator *M*, we open the non-causal path $A \rightarrow M \leftarrow U \rightarrow Y$ in the graphs in Figure 3.1, which then become mixed and confused with the direct causal path $A \rightarrow Y$. If opening the non-causal path is inevitable, one solution can be trying to offset the opened non-causal path by creating another equivalent non-causal path. Study 1 and Study 2 of this dissertation have shown that the underlying bias-removing mechanism of the standard difference-in-differences (i.e., gain score methods) and the fixed effects approach is bias-offsetting by creating another non-causal path(s). In Figure 3.2, we add a variable *D* called the *difference score* (or the gain or change score), which is calculated by $D = Y - P$. The dashed arrows represent relationships that were created by researchers, as opposed to by nature, through the implementation of a specific analytic model. Solid arrows represent causal relationships created by nature, implying they cannot be altered by researchers.

Given the graph, again assuming linear systems, the direct effect of *A* on *Y*, τ, is identifiable despite *U*, if $\gamma_Y = \gamma_P$ holds. This equality states that the unmeasured mediator-outcome confounder *U* directly affects the primary outcome *Y* and its pretest *P to the same extent*. We shall refer to this equality as the *equivalent mediator-outcome confounding assumption*. Under this assumption, one can correctly identify the direct causal effect. Note that conditioning on *M* will still create the collider bias, that is, the path $A \rightarrow M \leftarrow U \rightarrow Y$ or the bias term in Equation (3.2). But, the pretest plays a role in neutralizing this collider bias or path. In the graph, conditioning on *M* leaves the following paths open between *A* and *D* (not *Y*):

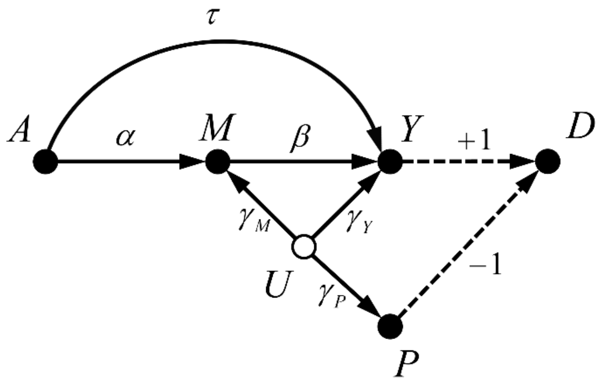

FIGURE 3.2. *Causal structure merging data-generating model (solid arrows) and analytic model (dashed arrows) using difference-in-differences in mediation analysis (DiDiM).*

(i) $A \to Y \to D$;

(ii) $A \to M \leftarrow U \to Y \to D$;

(iii) $A \to M \leftarrow U \to P \to D$.

Path (i) corresponds to the direct effect τ because the last causal link of the relationship $Y \to D$ of the path is indeed an identity transformation. The product of structural path coefficients of the path is $\tau \times (+1) = \tau$. Path (ii) corresponds to the original unmeasured mediator-outcome confounding, which was $A \to M \leftarrow U \to Y$ (without D) in Figure 3.1A. Again, these two paths are equivalent because the link $Y \to D$ is an identity transformation. However, note that by using the difference score D, we open another non-causal path—path (iii). But, path (ii) and path (iii) offset each other if $\gamma_Y = \gamma_P$. This is because the products of path coefficients of each of the two paths are identical in terms of magnitude but have different signs because of the link $P \to D$, where the path coefficient is -1. That is, the mediator-outcome confounding bias occurs due to conditioning on the mediator, but we create an equivalent confounding bias by using the pretest, and finally difference out all the biases.

This is algebraically verified. Let's regress the difference score *D* on the treatment *A* and the mediator *M*. The population partial regression coefficient of *A* is given by

$$b_{DA|M} = \tau - \frac{\alpha\gamma_M(\gamma_Y - \gamma_P)}{Var(M) - Var(A)\alpha^2} . \tag{3.5}$$

Again, it consists of the true direct effect and a bias term, but the bias term disappears if the equivalent mediator-outcome confounding assumption holds ($\gamma_Y = \gamma_P$). Note that the numerator of the second term of Equation (3.5) can be re-written as $\alpha\gamma_M\gamma_Y - \alpha\gamma_M\gamma_P$. The former $\alpha\gamma_M\gamma_Y$ corresponds to the product of path coefficients of path (ii), and the latter $-\alpha\gamma_M\gamma_P$ corresponds to that of path (iii) and they cancel each other out—again, the denominator or scaling factor is not represented by graphs.

We emphasize the different role of pretest measures in the DiDiM approach. Standard mediation analysis views the pretest as a proxy of the unmeasured mediator-outcome confounding variable and tries to reduce the confounding bias relying on the resemblance between the pretest and the confounding variable, expressed with the reliability in Equation (3.3). In contrast, the DiDiM views the pretest as a counterfactual outcome if the direct and indirect causal impacts of the treatment would have not been given. The DiDiM uses this pretest to quantify the extent of the confounding bias. In fact, these conflicting views on the role of pretests are what Lord (1967) raised earlier and what are now known as Lord's paradox (i.e., gain score method vs. ANCOVA). We extend Lord's setup to the mediation context for resolving the unmeasured mediator-outcome confounding problem. The basic features of the gain score method (or the difference-in-differences or fixed effects models) can also be applied to the

DiDiM. For example, DiDiM estimators hold the same three advantages as gain score estimators insofar as both estimators are robust against i) the unreliability of the pretest, ii) bias amplification, and iii) collider bias due to correlated error terms of P and Y (see Study 1 of this dissertation). All three threats will distort the standard mediation estimators which simply control for the mediator and the pretest. Below, we identify another advantage of DiDiM estimators which has never been discussed in the literature.

"Difference-in-Differences in Mediation Analysis" in Non-Experimental Studies

In order to focus on unmeasured mediator-outcome confounding, so far, we have assumed that the treatment assignment was randomized and thus both no treatment-mediator and no treatment-outcome confounding assumptions hold. Randomizing the treatment or, equivalently, establishing the conditional independence assumption with respect to the treatment conditional on some pre-treatment covariates (i.e., the first ignorability of the sequential ignorability assumption; see Imai et al, 2010b) is a basic requirement for causal identification of mediation effects. If the treatment itself is confounded (with the mediator or outcome), causal mediation effects are not generally identified regardless of the unmeasured mediator-outcome confounding.

However, this rationale only holds with the standard mediation approach. Interestingly, the DiDiM approach allows us to identify direct effects even though the treatment is not randomized and thus we cannot identify total effects and indirect effects. In Figure 3.3, we allow the arrow $U \rightarrow A$, implying that the unmeasured variable U confounds not only the mediator-outcome relationship ($M \leftarrow U \rightarrow Y$) but also the treatment-mediator ($A \leftarrow U \rightarrow M$) and the

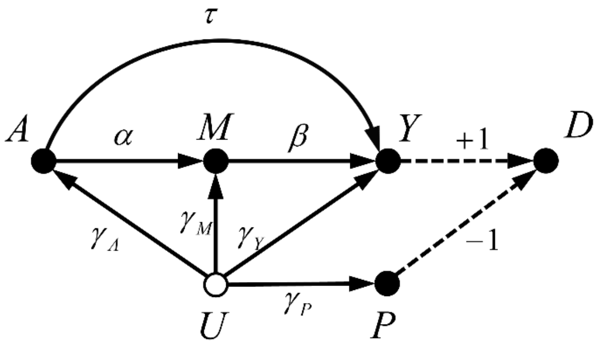

FIGURE 3.3. *Causal structure merging data-generating model (solid arrows) and analytic model (dashed arrows) using difference-in-differences in mediation analysis (DiDiM) when the treatment A is not randomized ($U \rightarrow A$ is added to Figure 3.2).*

treatment-outcome ($A \leftarrow U \rightarrow Y$) relationships.[6] According to the conventional criterion, we cannot identify the total effect but also cannot decompose the effect into the direct and indirect effects because all three confounding assumptions—assumptions about treatment-outcome, treatment-mediator, and mediator-outcome relationships—are violated. However, the DiDiM allows for the identification of the direct effect τ even in this case. In Figure 3.3, conditional on *M*, all the open paths between *A* and *D* are

[6] Strictly speaking, allowing the arrow $U \rightarrow A$ changes the structure in Figure 3.3 to a nonlinear system when *A* is binary. Then, the parameter γ_A, and any analytic formulae including it such as Equation (3.6), may be misleading. Although it is possible to derive a closed-form solution for the direct effect estimate from the structure in Figure 3.3 when *A* is binary, the resulting formula will be complicated without adding any substantive changes. In fact, whether the relationship between *U* and *A* is linear or nonlinear does not affect the mechanism of the DiDiM method. Any association via the relationship will be offset anyway. For ease of exposition, here we shall treat the parameter γ_A as any other parameters in this linear system. Readers interested in this issue as well as the possible accurate formulae for binary treatment may refer to Ding and Miratrix (2015) or Steiner and Kim (2016).

(i) $A \rightarrow Y \rightarrow D$;

(ii) $A \rightarrow M \leftarrow U \rightarrow Y \rightarrow D$;

(iii) $A \rightarrow M \leftarrow U \rightarrow P \rightarrow D$;

(iv) $A \leftarrow U \rightarrow Y \rightarrow D$;

(v) $A \leftarrow U \rightarrow P \rightarrow D$.

Note that path (i) corresponds to the causal direct effect while the other four paths are non-causal and represent confounding. If the equivalent mediator-outcome confounding assumption, $\gamma_Y = \gamma_P$, holds, then not only path (ii) and path (iii) but also path (iv) and path (v) offset each other; therefore, the direct effect of *A* on *D* ($A \rightarrow Y \rightarrow D$), which equals the direct effect of *A* on *Y* ($A \rightarrow Y$), can be identified using DiDiM. In fact, the offsetting of path (iv) and path (v) is what happens in the standard difference-in-differences method or gain score method, in which no mediators appear (Study 1 and Study 2 of this dissertation).

This can be algebraically shown. Given the graph in Figure 3.3, the population partial regression coefficient for *A* of the regression model of *D* on *A* and *M* is given by

$$\begin{aligned} b_{DA|M} &= \tau - \frac{\alpha\gamma_M(\gamma_Y - \gamma_P)Var(A)}{Var(A)Var(M) - \{Var(A)\alpha + \gamma_A\gamma_M\}^2} \\ &\quad + \frac{\gamma_A(\gamma_Y - \gamma_P)\{Var(M) - Var(A)\alpha^2 - \gamma_M^2 - \alpha\gamma_A\gamma_M\}}{Var(A)Var(M) - \{Var(A)\alpha + \gamma_A\gamma_M\}^2}. \end{aligned} \tag{3.6}$$

The second subtractive term in Equation (3.6), analogous to the bias term in Equation (3.5), corresponds to the offset of path (ii) and path (iii) because the product of path coefficients of path (ii) is $\alpha\gamma_M\gamma_Y$ and that of path (iii) is $-\alpha\gamma_M\gamma_P$. The third additive term corresponds to the offset of path (iv) and path (v) because the product of path coefficients of path (iv) is $\gamma_A\gamma_Y$ and that of path (v) is $-\gamma_A\gamma_P$. Under $\gamma_Y = \gamma_P$, both bias terms become zero and the partial regression coefficient becomes the causal direct effect τ.

Nonetheless, neither the standard mediation approach nor the DiDiM approach identify total effects or indirect effects. The reason why the DiDiM fails to identify total effects in particular is due to the non-causal path, $A \leftarrow U \rightarrow M \rightarrow Y \rightarrow D$, which cannot be offset under $\gamma_Y = \gamma_P$. This is an intriguing case where one may identify causal direct effects even though he or she cannot identify total effects and indirect effects. This case, which has never been discussed in the mediation literature, happens because of the unique bias-removing mechanism of the DiDiM approach, and thus this case highlights the fundamental difference between the standard mediation approach and the DiDiM approach.

Our intention here is not to say that the DiDiM is always able to eliminate all types of confounding biases and thus does not require any confounding assumptions (at least, to identify direct effects). In fact, it is easy to present a case where the DiDiM cannot eliminate treatment-mediator or treatment-outcome confounding and thus fails to identify causal direct effects when the treatment is not randomized. For example, if there is an unmeasured confounder between the treatment and outcome that is *independent* of the unmeasured mediator-outcome confounding variable (e.g., *U* in Figure 3.3), then the treatment-outcome confounding bias cannot be eliminated. Nonetheless, in observational studies, the DiDiM approach may help to eliminate some of the treatment-outcome confounding bias if the source of the confounding is correlated with the unmeasured mediator-outcome confounding variables. In this study’s illustration, we will show empirical evidence that the DiDiM outperforms the standard mediation analyses especially when the treatment is not randomized.

ADJUSTING "DIFFERENCE-IN-DIFFERENCES IN MEDIATION ANALYSIS" ESTIMATORS

The key assumption for causal identification using DiDiM is the equivalent mediator-outcome confounding, expressed with $\gamma_Y = \gamma_P$. If this is violated, the resulting DiDiM estimators will be biased. However, in practice, justifying this assumption is challenging. Basically, it requires that the pretest and the outcome should be measured on the same scale (Sofer, Richardson, Colicino, Schwartz, & Tchetgen Tchetgen, 2016). Advanced measurement techniques using item response theory for test equating, scaling, or calibrating may help to improve this scale-invariance. However, even though a test-administer may develop a perfectly equivalent test, subjects themselves might change over time and thus the equivalent mediator-outcome confounding assumption can be violated. Although we use the DiDiM to relax the no unmeasured confounding assumption, it also requires its own assumption, which, in practice, may not be easily met either.

We present a strategy to adjust DiDiM estimators for the extent of the violation of the equivalent mediator-outcome confounding assumption in linear systems. Miao and Tchetgen Tchetgen (2018) and Shi, Miao, and Tchetgen Tchetgen (2018) proposed a method, referred to as *double negative controls*, that combines negative outcome controls and negative treatment controls to correct an unmeasured treatment-outcome confounding bias in non-mediation studies. We use a similar method to correct unmeasured mediator-outcome confounding in the mediation context. This method requires an additional variable, which we shall refer to as the *compass variable*.[7] In Figure 3.4A, the variable *C* is affected by the unmeasured mediator-outcome

[7] Our compass variables correspond to the negative treatment controls used by Miao and Tchetgen Tchetgen (2018) and Shi et al. (2018). While their terminology comes from the convention in epidemiology (i.e., negative controls), the term "compass variable" emphasizes its specific role for adjusting DiDiM estimators. As in geometric composition, the compass variable

confounding variable *U*. We use this compass variable *C* in order to quantify the extent of the discrepancy between the two parameters γ_Y and γ_P. One may wonder what the difference between the compass variable *C* and the pretest *P* is. In Figure 3.4A, both are affected by *U* but one of them is used as a compass while the other is used as a pretest to compute the difference score. The difference is illustrated in Figure 3.4B where the compass *C* is affected by both *A* and *M*. In this case, *C* can still serve as a compass variable but cannot be used as a pretest. Thus, compass variables require a weaker assumption than pretests. Like pretests, they are associated with the pretest and outcome only through the unmeasured mediator-outcome confounding variable, but unlike pretests, they may be associated with the treatment and the mediator (see Study 1 of this dissertation, for insight into why pretests should not be associated with the treatment in non-mediation studies; also Imai & Kim, 2017). Formally, the compass variable should be independent of the pretest and the outcome conditional on the unmeasured confounding variable, the treatment, and the mediator, $C \coprod (P, Y) \,|\, U, A, M$.[8] This conditional independence holds in Figure 3.4A and also in Figure 3.4B where *C* is affected by *A* and *M*.

Having a valid compass variable, one specifies two regression models. First, the outcome *Y* is regressed on *C, A*, and *M*. The regression coefficient of the compass variable *C* then represents the association transmitted only via the path $C \leftarrow U \rightarrow Y$ (all other paths are blocked by conditioning on *A* and *M* in the regression model). The source of this association is the product of the two path coefficients along with the path, $\gamma_C \gamma_Y$. Due to conditioning on *A* and *M*,

is used, metaphorically, to measure "relative distance" between the compass-pretest association and the compass-outcome association although the compass itself has no absolute scale.

[8] Therefore, in principle, the compass variable may causally affect the unmeasured confounding variable *U*, $C \rightarrow U$.

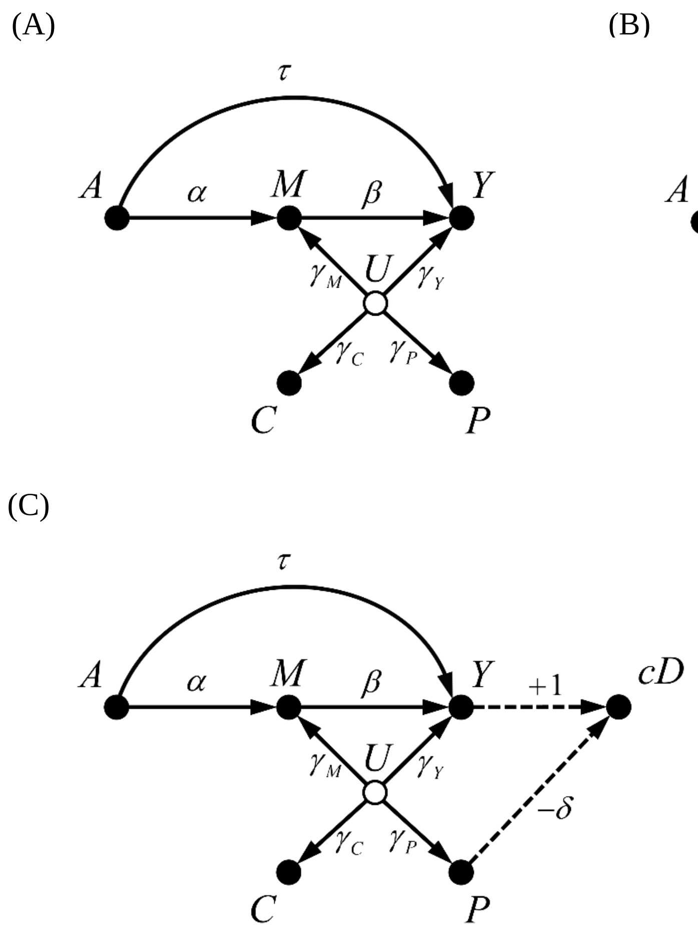

FIGURE 3.4. *Causal structures with the compass variable C. (A) C is added to the graph in Figure 1b. (B) Two arrows A → C and M → C are added to (A). (C) Corrected difference score cD is added to (A) by* $cD = Y - \delta P$.

resulting in partialling out some variation in *C*, the exact regression coefficient for *C* will be expressed as $b_{YC|AM} = \gamma_C \gamma_Y H$, where *H* is some multiplicative factor. Second, similarly, the pretest *P* is regressed on *C*, *A*, and *M*. The regression coefficient for *C* then represents the association transmitted via the path *C* ← *U* → *P*, and it is expressed as $b_{PC|AM} = \gamma_C \gamma_P H$. Note that the multiplicative factor *H* is identical because regressors are the same in both regression models. Then, the ratio δ of the two regression coefficients can quantify the discrepancy between γ_Y and γ_P : $\delta = \frac{b_{YC|AM}}{b_{PC|AM}} = \frac{\gamma_C \gamma_Y H}{\gamma_C \gamma_P H} = \frac{\gamma_Y}{\gamma_P}$. Any deviation from one indicates the violation of the equivalent mediator-outcome confounding assumption.

The coefficient ratio δ is used to compute the corrected difference score $cD = Y - \delta P$, instead of the original difference score $D = Y - P$. This analytic model is represented in Figure 3.4C using dashed arrows. The open paths between *A* and *cD* conditional on *M* are

(i) $A \rightarrow Y \rightarrow cD$;

(ii) $A \rightarrow M \leftarrow U \rightarrow Y \rightarrow cD$;

(iii) $A \rightarrow M \leftarrow U \rightarrow P \rightarrow cD$.

Note that the association transmitted via path (iii) is no longer the same as that in Figure 3.2 because the final link $P \rightarrow cD$ is not $\times(-1)$ but rather $\times(-\delta)$. The exact association via path (iii) can be expressed as $\alpha\gamma_M\gamma_P H' \times (-\delta)$, where H' denotes some multiplicative factor. As the association via path (ii) will be a form of $\alpha\gamma_M\gamma_Y H' \times (+1)$, the sum of the two associations via path (ii) and path (iii) becomes zero: $\alpha\gamma_M\gamma_P H' \times (-\delta) + \alpha\gamma_M\gamma_Y H' \times (+1) = -\alpha\gamma_M\gamma_Y H' + \alpha\gamma_M\gamma_Y H' = 0$, despite $\gamma_Y \neq \gamma_P$. Thus, path (ii) and path (iii) offset each other, and the overall association between *A* and *cD* is only via path (i), which represents the causal direct effect of *A* on *Y* because $\tau \times (+1) = \tau$.

In practice as instrumental variables, finding proper compass variables that satisfy the conditional independence $C \coprod (P, Y) \mid U, A, M$ might be challenging. Also, the adjustment using the coefficient ratio δ may be highly sensitive to a slight difference between two parameters γ_Y

and γ_P . More research is needed to further develop the adjustment of DiDiM using compass variables. In the illustration section that follows, we will provide positive empirical evidence to support such correction.

ILLUSTRATION

Benchmarks for Mediation Effects from Real Data

We illustrate the DiDiM approach using real data sets. Although it has a much stronger implication for practice, using real data, instead of simulated data, has one severe disadvantage: A valid benchmark is typically unknown to researchers. In our context, this means that we do not know the true (in)direct effects in real data sets that will be compared with the estimated effects by the proposed approach to evaluate its performance. To circumvent this challenge, we have developed a strategy that is useful for our illustration as well as potentially useful for other mediation studies. We shall refer to this as a *parallel-outcome design*. Consider a study where a treatment is randomized, and it affects two outcomes that cannot causally affect each other. For example, when a student takes his or her final exams in multiple subjects, the final score of one subject, for example, the math score, cannot causally affect the final score for the other subject, for example, the reading score (and vice versa). This is because the two scores are measured almost at the same time (same day or same week), and the scores, which are frozen into the answer sheets once submitted, are not typically known to students until all the exams are completed. Although in practice this may occasionally be violated, we believe such no causal relationships between two parallel outcome measures likely hold in many regular evaluation settings—especially, in the data sets we discuss below.

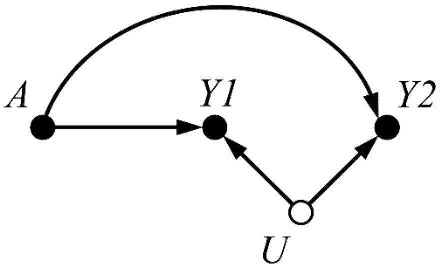

FIGURE 3.5. *Causal structure of a parallel-outcome design. Two outcomes Y1 and Y2 do not causally affect each other but are affected by the treatment A and their unmeasured common cause U.*

In Figure 3.5, we describe the causal structure of a parallel-outcome design where two outcomes *Y1* and *Y2* do not causally affect each other but they are affected by the treatment *A* and the unmeasured common cause *U*. In the graph, we view one outcome, for example, *Y1*, as a *pseudo-mediator*. For example, the reading score is a mediator, and the math score is a primary outcome. As we rule out any causal relationship between the two scores, the reading score is a mediator that does *not* convey any treatment effect via itself. This means that the true indirect or mediated effect via the reading score (i.e., pseudo-mediator) must be *zero*; consequently, the true total effect must equal the true direct effect. Note that the two outcomes can be and, indeed, likely are correlated, as represented by the path *Y1* ← *U* → *Y2*. For example, the math score and the reading score are correlated because of a student's ability, socioeconomic status, test anxiety, etc. All these factors are the source that creates the mediator-outcome confounding bias, with respect to the pseudo-mediator. In Figure 3.5, simply conditioning on *Y1* opens the non-causal or collider path,

$$A \rightarrow Y1 \leftarrow U \rightarrow Y2,$$

and thus, will produce a non-zero indirect effect estimate. However, if a mediation approach does not suffer from the mediator-outcome confounding bias or at least substantially mitigate this bias, the estimated indirect effect of A on $Y2$ via $Y1$ should be close to zero, and equivalently, the estimated direct effect of A on $Y2$ should be close to the total effect of A on $Y2$ under the parallel-outcome design.

Shadish, Clark, and Steiner's (2008) RCT and Non-RCT Data Sets[9]

Shadish, Clark, and Steiner (2008) investigated the effect of taking a math training session (compared to a vocabulary training session) on two outcomes, math and vocabulary scores (also their pretest scores that were measured before the training session). In their lab setting, it is implausible that any of the two scores determines the other score, because 30 vocabulary items and 20 math items are provided to subjects on a single test. We believe that the data-generating structure of Shadish et al.'s (2008) data fits the parallel-outcome design depicted in Figure 3.5. Therefore, we certainly know that the causal indirect effect of the math training on the math score (outcome) via the vocabulary score (pseudo-mediator) is zero. Since the outcome and the pseudo-mediator are parallel, we can switch them such that we also know that the causal indirect effect of the math training on the vocabulary score (outcome) via the math score (pseudo-mediator) is zero.

Furthermore, one of the design features used by Shadish et al. (2008) allows us to investigate how the DiDiM can eliminate confounding bias in a non-experimental setting. In

[9] We thank M. H. Clark for providing access to the data sets for the illustration.

their study, Shadish et al. (2008) constructed two data sets, RCT data and non-RCT data, by randomly assigning subjects to either data condition. The subjects who were assigned to the RCT data condition (*n* = 235) were further randomly assigned to either a math (*n* = 119) or vocabulary (*n* = 116) training session while those who were assigned to the non-RCT data condition (*n* = 210) self-selected their training session (n = 79 for math training; *n* = 131 for vocabulary training). Although the treatment is not randomized in the non-RCT data condition, the true causal effects can be inferred from the effects in the RCT data in this within-study comparison (Cook, Shadish, & Wong, 2008). Therefore, we know that in the non-RCT data condition, the causal indirect effect of the math training on the math (or vocabulary) score via the vocabulary (or math) score is also zero.

Using Shadish et al.'s (2008) data sets, we create four analysis conditions: 2 (data condition: RCT vs. non-RCT) × 2 (outcome-mediator combination: math-vocabulary vs. vocabulary-math). In each analysis condition, we estimate the causal direct effect of the math training using four different methods: i) *controlling for M*, ii) *controlling for M & P*, iii) *DiDiM*, and iv) *adjusted DiDiM*. The first two methods are the standard mediation approach—simply controlling for the mediator with or without the pretest. The last two methods are our proposed approach—the simple DiDiM and its correction using a compass variable. Subjects' ACT or SAT college admission score is used as our compass variable (subjects were undergraduate students) because we believe that the conditional independence $C \amalg (P, Y) \mid U, A, M$ likely holds with ACT scores.[10]

[10] Our rationale was that since ACT scores were measured before subjects entered college, it would be less likely that they are directly associated with both math and vocabulary outcomes not via the mediator-outcome confounding variable (e.g., *U* in Figure 3.5). Of course, this is our subjective judgement and may be incorrect. However, in the analyses, we observed the

TABLE 3.1.

Estimated direct effects of the math training session on the math and vocabulary scores via each of the pseudo-mediators

	Outcome = Math;
	Pseudo-Mediator = Vocabulary
			Outcome = Vocabulary;
			Pseudo-mediator = Math	

	Coef	Bootstrap 95% CI	Coef	Bootstrap 95% CI
Direct Effect (benchmark)	4.11	[3.28, 4.93]	−8.28	[−9.01, −7.50]
1. RCT data				
Controlling for M	6.14	[4.90, 7.33]	−9.10	[−9.97, −8.15]
Controlling for M & P	5.49	[4.20, 6.67]	−8.81	[−9.65, −7.89]
DiDiM	4.11	[2.65, 5.63]	−8.19	[−9.56, −6.97]
Adjusted DiDiM	3.70	[1.61, 5.33]	−8.66	[−9.60, −7.70]
2. Non-RCT data				
Controlling for M	7.60	[6.21, 9.00]	−10.22	[−11.21, −9.23]
Controlling for M & P	5.83	[4.51, 7.24]	−9.03	[−10.01, −8.08]
DiDiM	4.71	[3.53, 5.99]	−6.83	[−8.21, −5.41]
Adjusted DiDiM	3.75	[1.88, 5.42]	−8.60	[−9.65, −7.53]

Note. The bench mark direct effects are obtained from the RCT data, controlling for both math and vocabulary pretest scores.

Results

Table 3.1 summarizes the results. Let us first focus on the condition of the math outcome (mediator: vocabulary) in the left column of Table 3.1. In RCT data, controlling for the pseudo-mediator (i.e., “Controlling for M”) estimates the direct treatment effect as 6.14 even though the benchmark direct effect is 4.11. Thus, the standard method overestimates the direct effect because of unmeasured mediator-outcome confounding. This overestimation is indeed expected. The math training (*A*) negatively affects the pseudo-mediator (*Y1*), vocabulary score, because the

correction that using ACT scores substantially improves the causal mediation estimation using DiDiM.

control condition is taking the vocabulary training session, and the two outcomes *Y1* and *Y2* are likely positively correlated with each other. Therefore,

$$\text{sgn}(A \rightarrow Y1 \leftarrow U \rightarrow Y2) = \text{sgn}(A \rightarrow Y1) \times \text{sgn}(Y1 \leftarrow U \rightarrow Y2) < 0 .$$

But, as this path becomes open due to conditioning on the collider, the sign of the overall association via this collider path is *reversed* in linear systems. Therefore, the overall mediator-outcome confounding bias via the path becomes positive, resulting in an overestimation of the direct effect. This overestimated direct effect decreases to 5.49 if one additionally controls for the pretest of the math outcome (i.e., "Controlling for M & P"), but it continues to suffer from the bias because it is still far greater than the benchmark 4.11. However, the DiDiM method corrects this overestimation. The point estimate by the DiDiM is 4.11, which is very close to the benchmark (indeed, they are identical up to two decimal points). We also apply the correction method using the compass variable, ACT scores (i.e., "Adjusted DiDiM"). It slightly underestimates the direct effect as 3.70, but it is still close to the benchmark (4.11) and is obviously far less biased than the estimates by two standard mediation methods.

In non-RCT data, we found that the four different methods have the same relative bias-removing potential. See the bottom of Table 3.1 in the same math outcome column. The method "Controlling for M" makes more bias in the non-RCT data condition than it does in the RCT data condition because the treatment is not randomized. In addition to the mediator-outcome confounding bias, the treatment-mediator and the treatment-outcome confounding biases contribute to making the largest biased estimate: 7.60. Although the additional controlling for the pretest reduces the bias to 5.83, it remains greater than the two DiDiM methods—the unadjusted (4.71) and adjusted (3.75) ones. Precisely, in the non-RCT data condition, the adjusted DiDiM

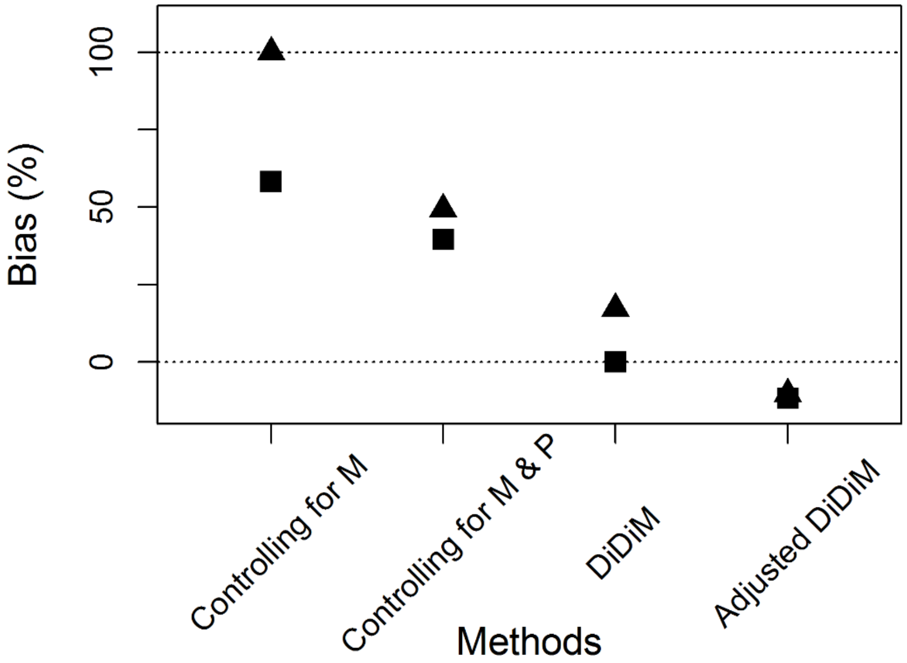

FIGURE 3.6. *Remaining bias in direct effects estimates of the treatment on math score (mediator: vocabulary score) by four mediation analysis methods. Triangles represent the results using non-RCT data, while squares represent the results using RCT data.*

produced the least biased estimate (the benchmark is 4.11). All the results are summarized and displayed in Figure 3.6, where squares represent the RCT data results and triangles represent the non-RCT data results. Overall, the standard mediation methods, "Controlling for M" and "Controlling for M & P," make more bias than the two DiDiM methods. The correction of the DiDiM using a compass variable was especially effective in eliminating confounding bias when the treatment was not randomized.

The results of switching the mediator and the outcome, that is, the vocabulary outcome condition (mediator: math), are presented in the right column of Table 3.1 and in Figure 3.7. Overall, the bias pattern was similar to the extent that the two standard mediation methods,

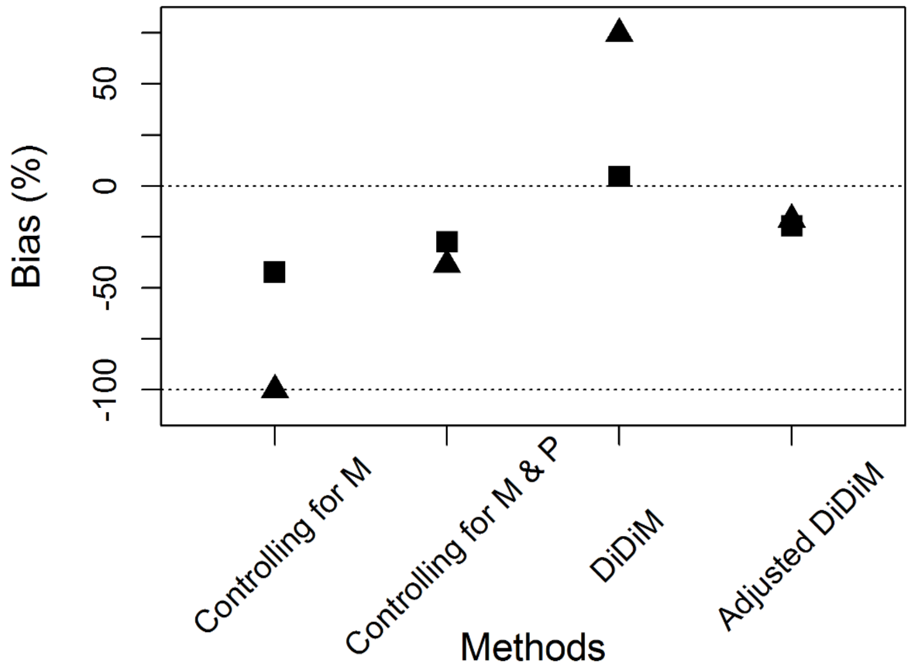

FIGURE 3.7. *Remaining bias in direct effects estimates of the treatment on vocabulary score (mediator: math score) by four mediation analysis methods. Triangles represent the results using non-RCT data, while squares represent the results using RCT data.*

simply controlling for the mediator with or without the pretest, make more bias than the two DiDiM methods. We shall not repeat the similar results but focus on one intriguing pattern we found. The unadjusted or simple DiDiM estimate using the non-RCT data was −6.83, which overestimates the true direct effect of −8.28. As we briefly discussed before, this misleading result suggests that the DiDiM may fail to eliminate all confounding bias when the treatment is not randomized, possibly because of an uncorrelated treatment-outcome confounding variable. However, interestingly, the adjusted DiDiM using the compass variable corrected this substantial bias and produced a far less biased effect estimate, one that was especially lower than the standard method of “Controlling for M & P” with the RCT data. This suggests that the use of a compass variable may also correct not only the mediator-outcome confounding but also the treatment-outcome confounding. But, this is a single empirical finding, and more research is needed to clarify the role of compass variables in adjusting DiDiM estimators in observational studies.

DISCUSSION

In this paper, we proposed a new mediation approach to circumvent the unmeasured mediator-outcome confounding problem. The key idea of our methods is the use of difference-in-differences to offset such confounding. With a proper pretest that establishes the equivalent mediator-outcome confounding assumption, we can eliminate the bias and identify mediation effects even though we do not measure all confounding variables. In the illustration, we provided empirical evidence to support the use of the DiDiM approach. If the assumption is severely violated or the treatment is not randomized, one may consider applying the correction method using compass variables. Again, we provided empirical evidence to support the use of this kind of correction.

Implementing the proposed approaches, the DiDiM and its correction, does not require any special software or skills, so practitioners can easily apply these methods to their mediational research questions if a pretest measure of the outcome is available. Without extra cost, they can explore how the results of the standard mediation methods and the DiDiM methods differ and evaluate which causal assumption (either no mediator-outcome confounding or equivalent mediator-outcome confounding) is more plausible in their studies. This theoretical consideration can be facilitated by the use of DiDiM methods, and we believe this will improve causal mediation analysis in practice, which currently almost exclusively relies on the standard methods of using covariate adjustment.

Obviously, the proposed methods need to be further developed and extended. In this paper, our discussion is based on linear systems, and we did not reflect the various causal mediation effects defined with counterfactual or potential outcomes (e.g., Imai et al., 2010a, 2010b; Pearl, 2014, VanderWeele, 2010). More than direct and indirect effects, recently causal inference researchers have started defining and using *controlled* direct effects and *natural* direct and indirect effects for causal mediation effects (Pearl, 2001). More investigation is needed to understand how and under which assumptions each of these effects can be nonparametrically identified by the DiDiM. These assumptions may or may not be similar to the conditions for nonparametric identification of controlled direct effects and natural (in)direct effects using the standard mediation approach based on covariate adjustment. Finally, the correction method using compass variables should be further studied. How the method works in observational studies, how to assess a good compass variable, and which design elements allow the secure preservation of a compass variable are all questions that should be addressed for the general use of this method.

CONCLUSION

This dissertation investigates the role of pretest measures of the outcome in causal inference. Pretests can be used to *offset* not only the treatment-outcome cofounding bias (i.e., Study 1 and Study 2) but also the mediator-outcome cofounding bias in mediation analysis (i.e., Study 3). The bias-offsetting mechanism is attractive because it allows us to identify causal effects even though the confounding variables remain unmeasured and thus it is infeasible to condition on, or control for, the variables to block non-causal paths. This is a special role that pretests can effectively play but that other regular covariates generally cannot. The key identification assumption for the offsetting mechanism, referred to as the common trend or the fixed effects assumption in Study 1 and Study 2 and as the equivalent mediator-confounding assumption in Study 3, is more plausible with a pretest rather than any other variables.

Proposed graphical models in many figures of the three studies are not supplementary materials but indeed the main results of this dissertation. This dissertation proposes for the first time causal graphical representations for the gain score model, the deviation score model, and the random effects approach (i.e., the quasi-deviation score model). In the literature, graphical models have been frequently used to describe how data are generated (e.g., Elwert, 2013; Morgan & Winship, 2015; Pearl, 2009; Rohrer, 2018; Steiner et al., 2017). As the dissertation shows, graphical models can also be used to represent analytic models or methods, and they are surprisingly helpful in understanding the power and the limitation of each of the analytic models for causal identification. As is seen in Study 3, the graphical representations for the random effects approach directly shows why the approach is inherently limited to being used for causal identification in non-experimental or observational studies.

The three studies in this dissertation collectively underscore the importance of having strong knowledge about data-generating models and analytic models. More precisely, for making a causal inference from data, selecting proper analytic models or methods should be guided by researchers' substantive knowledge about the data-generating process. Pretests are an excellent example of this point. Although many educational and psychological researchers have implicitly agreed that pretests are important, a formal theory of what a good pretest is and how we can use it for identifying causal effects (the main theme of this dissertation) has remained unclear in the literature. In fact, this is the key to the well-known Lord's paradox (Lord, 1967) discussed in Study 1. With a single pretest measure, two statisticians, each of whom applied a different analytic method, reach the opposite conclusions about the treatment effect. Somewhat ironically, the solution to the paradox can be given by the client who asked for the statistical consulting because he or she (but not the statisticians) knows well how the data were generated. The two statisticians should "consult" with the client to judge which conclusion is right. As this dissertation shows, graphical models will significantly facilitate the communication between the client and the statisticians.

REFERENCES

Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American statistical Association, 105*(490), 493-505.

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions.* Newbury Park, CA: Sage.

Allison, P. D. (1990). Change scores as dependent variables in regression analysis. *Sociological Methodology*, *20*(1), 93-114.

Allison, P. D. (1994). Using panel data to estimate the effects of events. *Sociological Methods & Research, 23*(2), 174-199.

Allison, P. D. (2009). *Fixed effects regression models* (Vol. 160). Thousand Oaks, CA: SAGE publications, Inc.

Ashenfelter, O. & Krueger, A. B. (1994). Estimates of the economic return to schooling from a new sample of twins. *American Economic Review, 84*(5), 1157-1173.

Baltagi, B. H. (2001). *Econometric analysis of panel data*. Chichester, UK: Wiley and Sons.

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*(6), 1173-1182.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1-48. doi:10.18637/jss.v067.i01.

Beck, N. (2011). Of fixed-effects and time-invariant variables. *Political Analysis, 19*(2), 119-122.

Bell, A., & Jones, K. (2015). Explaining fixed effects: Random effects modeling of time-series cross-sectional and panel data. *Political Science Research and Methods*, *3*(1), 133-153.

Bollen, K. A., & Brand, J. E. (2010). A general panel model with random and fixed effects: A structural equations approach. *Social Forces, 89*(1), 1-34.

Bollen, K. A., & Pearl, J. (2013). Eight myths about causality and structural equation models. In S. L. Morgan (Ed.), *Handbook of causal analysis for social research* (pp. 301–328). New York, NY: Springer.

Breusch, T., Ward, M. B., Nguyen, H. T. M., & Kompas, T. (2011a). On the fixed-effects vector decomposition. *Political Analysis*, *19*(2), 123-134.

Breusch, T., Ward, M. B., Nguyen, H. T. M., & Kompas, T. (2011b). FEVD: Just IV or just mistaken? *Political Analysis*, *19*(2), 165-169.

Campbell, D. T., & Erlebacher, A. (1970). How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education programs look harmful. In J. Hellmuth (Ed.), *The Disadvantaged Child: Vol. 3. Compensatory Education: A National Debate* (pp. 185-210). New York, NY: Bruner/Mazel.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 171-246). Chicago, IL: Rand McNally.

Caro, D. H. (2015). Causal mediation in educational research: An illustration using international assessment data. *Journal of Research on Educational Effectiveness*, *8*(4), 577-597.

Clark, T. S., & Linzer, D. A. (2015). Should I use fixed or random effects? *Political Science Research and Methods*, *3*(2), 399-408.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Boston, MA: Houghton Mifflin.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management, 27*(4), 724-750.

Cook, T. D., & Steiner, P. M. (2010). Case matching and the reduction of selection bias in quasi-experiments: The relative importance of pretest measures of outcome, of unreliable measurement, and of mode of data analysis. *Psychological Methods, 15*(1), 56-68.

Cronbach, L. J., & Furby, L. (1970). How we should measure "change": Or should we? *Psychological Bulletin, 74*(1), 68-80.

Darley, J. (2001, July 1). *The tradition of experimentalism in psychology*. Retrieved from https://www.psychologicalscience.org/observer/the-tradition-of-experimentalism-in-psychology

Dee, T. S. & Jacob, B. A. (2011). The impact of the No Child Left Behind Act on student achievement. *Journal of Policy Analysis and Management, 30*(3), 418–446.

Ding, P., & Miratrix, L. W. (2015). To adjust or not to adjust? Sensitivity analysis of M-bias and butterfly-bias. *Journal of Causal Inference*, *3*(1), 41-57.

Ding, P., & Vanderweele, T. J. (2016). Sharp sensitivity bounds for mediation under unmeasured mediator-outcome confounding. *Biometrika*, *103*(2), 483-490.

Dunn, G., & Bentall, R. (2007). Modelling treatment-effect heterogeneity in randomized controlled trials of complex interventions (psychological treatments). *Statistics in Medicine, 26*(26), 4719-4745.

Elwert, F. (2013). Graphical causal models. In: S. Morgan (ed.), *Handbook of Causal Analysis for Social Research*. Dodrecht: Springer.

Elwert, F., & Winship, C. (2014). Endogenous selection bias: The problem of conditioning on a collider variable. *Annual Review of Sociology, 40*, 31-53.

Festinger, L., & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *The Journal of Abnormal and Social Psychology*, *58*(2), 203-210.

Gelman A, & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York, NY: Cambridge University Press.

Glynn, A. (2012). The product and difference fallacies for indirect effects. *American Journal of Political Science, 56*(1), 257–269.

Greene, W. (2011). Fixed effects vector decomposition: A magical solution to the problem of time-invariant variables in fixed effects models? *Political Analysis, 19*(2), 135-146.

Gunasekara, F. I., Richardson, K., Carter, K., & Blakely, T. (2014). Fixed effects analysis of repeated measures data. *International Journal of Epidemiology, 43*(1), 264-269.

Halaby, C. (2004). Panel models in sociological research: Theory into practice. *Annual Review of Sociology, 30*, 507–544.

Hallberg, K., Cook, T. D., Steiner, P. M., & Clark, M. H. (2016). Pretest measures of the study outcome and the elimination of selection bias: Evidence from three within study comparisons. *Prevention Science*, 1-10.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica*, *46*(6), 1251-1271.

Hayes, A. F. (2009). Beyond Baron and Kenny: Statistical mediation analysis in the new millennium. *Communication Monographs, 76*(4), 408-420.

He, P., Wu, Z., Zhang, X. D., & Geng, Z. (2016). Identification of causal mediation models with an unobserved pre-treatment confounder. In H. He, P. Wu, & D.G. Chen (Eds.), *Statistical causal inferences and their applications in public health research* (pp. 241-262). Switzerland, Springer International Publishing.

Huber, E., & Stephens, J. D. (2000). Partisan governance, women's employment, and the social democratic service state. *American Sociological Review, 65*(3), 323-342.

Hsiao, C. (2003). *Analysis of panel data* (2nd ed.). New York, NY: Cambridge University Press.

Imai, K., Keele, L., & Tingley, D. (2010a). A general approach to causal mediation analysis. *Psychological Methods*, *15*(4), 309.

Imai, K., Keele, L., & Yamamoto, T. (2010b). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, *25*(1), 51-71.

Imai, K., & Kim, I. S. (2017). *When should we use linear fixed effects regression models for causal inference*. Princeton, NJ: Princeton University. Retrieved from https://imai.princeton.edu/research/files/FEmatch.pdf

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics*, *86*, 4–29.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction*. New York, NY: Cambridge University Press.

Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, *47*(1), 5-86.

Jamieson, J. (2004). Analysis of covariance (ANCOVA) with difference scores. *International Journal of Psychophysiology, 52*(3), 277-283.

Judd, C. M., & Kenny, D. A. (1981). Process analysis: Estimating mediation in evaluation research. *Evaluation Research, 5,* 602-619.

Kaplan, D. (2009). Causal inference in non-experimental educational policy research. In G. Sykes, B. Schneider, & D. N. Plank (Eds.), *Handbook on education policy research* (pp. 139–153). New York, NY: Taylor and Francis.

Kenny, D. A. (1975). Cross-lagged panel correlation: A test for spuriousness. *Psychological Bulletin, 82,* 887-903.

Laird, N. (1983). Further comparative analyses of pretest-posttest research designs. *The American Statistician, 37*(4), 329-330.

Lechner, M. (2011). The estimation of causal effects by difference-in-difference methods. *Foundations and Trends® in Econometrics, 4*(3), 165-224.

Lipsitch, M., Tchetgen Tchetgen, E., & Cohen, T. (2010). Negative controls: A tool for detecting confounding and bias in observational studies. *Epidemiology, 21*(3), 383-388.

Lockwood, J. R., & McCaffrey, D. F. (2007). Controlling for individual heterogeneity in longitudinal models, with applications to student achievement. *Electronic Journal of Statistics, 1,* 223-252.

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin, 68*(5), 304-305.

MacKinnon, D. (2008). *Introduction to statistical mediation analysis*. New York, NY: Erlbaum.

MacKinnon, D., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods, 7*(1), 83-104.

MacKinnon, D., Lockwood, C. M., Brown, C., Wang, W., & Hoffman, J. (2007). The intermediate endpoint effect in logistic and probit regression. *Clinical Trials, 4*, 499-513.

Miao, W., & Tchetgen Tchetgen, E. (2018, May). *Identifying causal effects with negative controls: Repair an invalid IV*. Paper presented at the Atlantic Causal Inference Conference, Pittsburgh, PA.

Maris, E. (1998). Covariance adjustment versus gain scores—revisited. *Psychological Methods, 3*(3), 309-327.

Morgan, S. L., & Winship, C. (2015). *Counterfactuals and causal inference: Methods and principles for social research* (2nd ed.). New York, NY: Cambridge University Press.

Myers, J. A., Rassen, J. A., Gagne, J. J., Huybrechts, K. F., Schneeweiss, S., Rothman, K. J., Joffe, M. M., & Glynn, R. J. (2011). Effects of adjusting for instrumental variables on bias and precision of effect estimates. *American Journal of Epidemiology*, *174,* 1213-1222.

Nichols, S. L., & Berliner, D. C. (2005). *The inevitable corruption of indicators and educators through high-stakes testing* (No. EPSL0503-101-EPRU). Tempe, AZ: Education Policy Studies Laboratory, Arizona State University. Retrieved July 1, 2018, from http://www.asu.edu/educ/epsl/EPRU/documents/EPSL-0503-101 -EPRU-exec.pdf

Nielsen, F., & Alderson, A. S. (1995). Income inequality, development, and dualism: Results from an unbalanced cross-national panel. *American Sociological Review*, *60*(5), 674-701.

Pearl, J. (1988). *Probabilistic inference in intelligent systems*. San Mateo, CA: Margan Kaufmann.

Pearl, J, (1993). Comment: Graphical models, causality, and intervention. *Statistical Science, 8*(3), 266-269.

Pearl, J. (2001). Direct and indirect effects. In J. Breese & D. Koller (Eds.), *Proceeding of the Seventeenth Conference on Uncertainty in Artificial Intelligence* (pp. 411-420). San Francisco, CA: Morgan Kaufmann.

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). New York, NY: Cambridge University Press.

Pearl, J. (2010). On a class of bias-amplifying variables that endanger effect estimates. In P. Grunwald, & P. Spirtes (Eds.), *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence* (pp. 425–432). Corvallis, OR: AUAI Press.

Pearl, J. (2011). Understanding bias amplification [Invited commentary]. *American Journal of Epidemiology, 174,* 1223-1227.

Pearl, J. (2013). Linear models: A useful “microscope” for causal analysis. *Journal of Causal Inference, 1*(1), 155-170.

Pearl, J. (2014). Interpretation and identification of causal mediation. *Psychological Methods, 19*(4), 459-481.

Pearl, J. (2016). Lord’s paradox revisited–(Oh Lord! Kumbaya!). *Journal of Causal Inference, 4*(2). Advance online publication. doi: 10.1515/jci-2016-0021

Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal inference in statistics: A primer*. New York, NY: Wiley.

Plümper, T., & Troeger, V. E. (2007). Efficient estimation of time-invariant and rarely changing variables in finite sample panel analyses with unit fixed effects. *Political Analysis, 15*(2), 124-139.

Plümper, T., & Troeger, V. E. (2011). Fixed-effects vector decomposition: properties, reliability, and instruments. *Political Analysis, 19*(2), 147-164.

Preacher, K. J., Rucker, D. D., & Hayes, A. F. (2007). Addressing moderated mediation hypotheses: Theory, methods, and prescriptions. *Multivariate Behavioral Research, 42*(1), 185-227.

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods, 40*(3), 879-891.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: SAGE publications, Inc.

Robins, J. M. (1987). A new approach to causal inference in mortality studies with a sustained exposure period—Application to control of the healthy worker survivor effect. *Mathematical Modelling, 7*, 1393-1512.

Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science, 1*(1), 27-42.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41-55.

Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association, 103*(484), 1334-1344.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton-Mifflin.

Shahar, E., & Shahar, D. J. (2012). Causal diagram and change variable. *Journal of Evaluation in Clinical Practice, 18*, 143-148.

Shi, X., Miao, W., & Tchetgen Tchetgen, E. (2018, May). *Multiply robust causal inference with double negative control adjustment for unmeasured confounding*. Paper presented at the Atlantic Causal Inference Conference, Pittsburgh, PA.

Small, D. S. (2012). *Mediation analysis without sequential ignorability: Using baseline covariates interacted with random assignment as instrumental variables*. Retrieved July 8, 2018, from https://arxiv.org/abs/1109.1070

Smolkowski, K. (2013, September 26). *Gain Score Analysis*. Retrieved December 15, 2017, from http://homes.ori.org/keiths/Tips/Stats_GainScores.html

Sobel, M. E. (2012). Does marriage boost men's wages?: Identification of treatment effects in fixed effects regression models for panel data. *Journal of the American Statistical Association, 107*(498), 521-529.

Sofer, T., Richardson, D. B., Colicino, E., Schwartz, J., & Tchetgen, E. J. T. (2016). On negative outcome control of unobserved confounding as a generalization of difference-in-differences. *Statistical Science, 31*(3), 348-361.

St.Clair, T. S., Cook, T. D., & Hallberg, K. (2014). Examining the internal validity and statistical precision of the comparative interrupted time series design by comparison with a randomized experiment. *American Journal of Evaluation, 35*(3), 311-327.

Steiner, P. M., Cook, T. D., & Shadish, W. R. (2011). On the importance of reliable covariate measurement in selection bias adjustments using propensity scores. *Journal of Educational and Behavioral Statistics*, *36*(2), 213-236.

Steiner, P. M., & Kim, Y. (2016). The mechanics of omitted variable bias: bias amplification and cancellation of offsetting biases. *Journal of Causal Inference, 4*(2). Advance online publication. doi: 10.1515/jci-2016-0009

Steiner, P. M., Kim, Y., Hall, C. E., & Su, D. (2017). Graphical models for quasi-experimental designs. *Sociological Methods & Research, 46*(2), 155-188.

Stuart, E. A., Huskamp, H. A., Duckworth, K., Simmons, J., Song, Z., Chernew, M. E., & Barry, C. L. (2014). Using propensity scores in difference-in-differences models to estimate the effects of a policy change. *Health Services and Outcomes Research Methodology, 14*(4), 166-182.

Ten Have, T. R., Joffe, M. M., Lynch, K. G., Brown, G. K., Maisto, S. A., & Beck, A. T. (2007). Causal mediation analyses with rank preserving models. *Biometrics, 63*(3), 926-934.

Ten Have, T. R., & Joffe, M. M. (2012). A review of causal estimation of effects in mediation analyses. *Statistical Methods in Medical Research, 21*(1), 77-107.

Thoemmes, F., & Mohan, K. (2015). Graphical representation of missing data problems. *Structural Equation Modeling: A Multidisciplinary Journal, 22*(4), 631-642.

Thomas, D. R., & Zumbo, B. D. (2012). Difference scores from the point of view of reliability and repeated-measures ANOVA in defense of difference scores for data analysis. *Educational and Psychological Measurement, 72*(1), 37-43.

Turkheimer, E., & Harden, K. P. (2014). Behavior genetic research methods: Testing quasi-causal hypotheses using multivariate twin data. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 159–187). New York, NY: Cambridge University Press.

Van Breukelen, G. J. (2006). ANCOVA versus change from baseline: More power in randomized studies, more bias in nonrandomized studies. *Journal of Clinical Epidemiology, 59*(9), 920-925.

VanderWeele, T. J. (2010). Bias formulas for sensitivity analysis for direct and indirect effects. *Epidemiology, 21*(4), 540-551.

VanderWeele, T. J. (2015). *Explanation in causal inference: Methods for mediation and interaction*. New York, NY: Oxford University Press.

VanderWeele, T. J. (2016). Mediation analysis: A practitioner's guide. *Annual Review of Public Health, 37*, 17-32.

Wichman, C. J., & Ferraro, P. J. (2017). A cautionary tale on using panel data estimators to measure program impacts. *Economics Letters, 151*, 82-90.

Wong, M., Cook, T. D., & Steiner, P. M. (2009). *No Child Left Behind: An interim evaluation of its effects on learning using two interrupted time series each with its own non-equivalent comparison series* (W-09–11). Evanston, IL: Institute for Policy Research.

Wong, V. C., Valentine, J. C., & Miller-Bains, K. (2017). Empirical performance of covariates in education observational studies. *Journal of Research on Educational Effectiveness, 10*(1), 207-236.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data* (2nd ed.). Cambridge, MA: The MIT Press.

Wooldridge, J. M. (2012). *Introductory econometrics: A modern approach* (5th ed.). South-Western Cengage Learning, Mason.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research, 20*(7), 557-585.

Zimmerman, D. W., & Williams, R. H. (1982). Gain scores in research can be highly reliable. *Journal of Educational Measurement, 19*(2), 149-154.

APPENDICES

Appendix A: Regression Estimator Formula in Study 1

The structural causal model corresponding to Figure 1.2D is given by $A = \varepsilon_A$, $Z = \alpha A + \varepsilon_Z$, $Y = \tau Z + \beta_2 A + \lambda_2 E + \varepsilon_Y$, $P = \beta_1 A + \lambda_1 E + \varepsilon_P$, and $G = Y - P$, where ε_A, ε_Z, ε_P, and ε_Y are the mutually independent random disturbance terms (omitted from the graph). Without loss of generality, we assume $Var(A) = Var(E) = 1$. Then, the partial regression coefficient of *Z*, $b_{YZ|P}$, of the regression of *Y* on *Z* and *P* can be written in terms of bivariate correlations as $b_{YZ|P} = \dfrac{\rho_{YZ} - \rho_{YP}\rho_{ZP}}{1-\rho_{ZP}^2} \times \dfrac{SD(Y)}{SD(Z)}$, where ρ_{AB} is the correlation coefficient between two random variables *A* and *B*. The correlation coefficients are given by

$$\rho_{YZ} = Cov(\tau Z + \beta_2 A + \lambda_2 E + \varepsilon_Y,\, \alpha A + \varepsilon_Z)/\{SD(Y)SD(Z)\} = \{Var(Z)\tau + \alpha\beta_2\}/\{SD(Y)SD(Z)\},$$

$$\rho_{YP} = Cov(\tau Z + \beta_2 A + \lambda_2 E + \varepsilon_Y,\, \beta_1 A + \lambda_1 E + \varepsilon_P)/\{SD(Y)SD(P)\}$$

$$= (\tau\alpha\beta_1 + \beta_1\beta_2 + \lambda_1\lambda_2)/\{SD(Y)SD(P)\},$$

$$\rho_{ZP} = Cov(\alpha A + \varepsilon_Z,\, \beta_1 A + \lambda_1 E + \varepsilon_P)/\{SD(Z)SD(P)\} = \alpha\beta_1/\{SD(Z)SD(P)\}.$$

Plugging the population correlations into the formula of $b_{YZ|P}$ above, we obtain Equation (1.6), $b_{YZ|P} = \tau + \dfrac{\alpha\beta_2(1-r)}{Var(Z) - \alpha^2 r} - \dfrac{\alpha\beta_1\lambda_1\lambda_2}{\{Var(Z) - \alpha^2 r\}Var(P)}$, where *r* is the reliability of the pretest *P*, $r = \beta_1^2/Var(P)$.

Because the structural causal model in Figure 1.2B is a restricted model of the structural causal model in Figure 1.2D, we obtain Equation (1.1) by setting $\lambda_2 = 0$:

$b_{YZ|P} = \tau + \frac{\alpha\beta_2(1-r)}{Var(Z) - \alpha^2 r}$. Note that the corresponding reliability $r = \frac{\beta_1^2}{\beta_1^2 + \lambda_1^2}$ is obtained by setting $Var(\varepsilon_P) = 0$ and treating *E* as *e*.

The structural causal model corresponding to Figure 1.2C is given by $A = \varepsilon_A$, $S = \varepsilon_S$, $Z = \alpha A + \alpha_S S + \varepsilon_Z$, $Y = \tau Z + \beta_2 A + \beta_S S + \varepsilon_Y$, $P = \beta_1 A + \varepsilon_P$, and $G = Y - P$. The correlation coefficients, based on this model, are given by

$$\begin{aligned}
\rho_{YZ} &= Cov(\tau Z + \beta_2 A + \beta_S S + \varepsilon_Y, \alpha A + \alpha_S S + \varepsilon_Z)/\{SD(Y)SD(Z)\} \\
&= \{Var(Z)\tau + \alpha\beta_2 + \alpha_S\beta_S\}/\{SD(Y)SD(Z)\}, \\
\rho_{YP} &= Cov(\tau Z + \beta_2 A + \beta_S S + \varepsilon_Y, \beta_1 A + \varepsilon_P)/\{SD(Y)SD(P)\} \\
&= (\tau\alpha\beta_1 + \beta_1\beta_2)/\{SD(Y)SD(P)\}, \\
\rho_{ZP} &= Cov(\alpha A + \alpha_S S + \varepsilon_Z, \beta_1 A + \varepsilon_P)/\{SD(Z)SD(P)\} = \alpha\beta_1/\{SD(Z)SD(P)\}.
\end{aligned}$$

Plugging them into $b_{YZ|P} = \frac{\rho_{YZ} - \rho_{YP}\rho_{ZP}}{1 - \rho_{ZP}^2} \times \frac{SD(Y)}{SD(Z)}$, we obtain Equation (1.3)

$$b_{YZ|P} = \tau + \frac{\alpha\beta_2(1-r)}{Var(Z) - \alpha^2 r} + \frac{\alpha_s\beta_s}{Var(Z) - \alpha^2 r}.$$

Relying on the same structural causal model, we can also obtain Equations (1.2), (1.4) and (1.5). First, Equation (1.4) is the regression coefficient for *Z* of the regression of *Y* on *Z*, b_{YZ}. Since the coefficient can be written as $b_{YZ} = Cov(Y, Z)/Var(Z)$ and using

$$Cov(Y, Z) = Cov(\tau Z + \beta_2 A + \beta_S S + \varepsilon_Y, \alpha A + \alpha_S S + \varepsilon_Z) = Var(Z)\tau + \alpha\beta_2 + \alpha_S\beta_S,$$

we obtain $b_{YZ} = \tau + \frac{\alpha\beta_2}{Var(Z)} + \frac{\alpha_S\beta_S}{Var(Z)}$. Similarly, for Equation (1.5), using

$$\begin{aligned}
Cov(G, Z) &= Cov(\tau Z + \beta_2 A + \beta_S S + \varepsilon_Y - \beta_1 A - \varepsilon_P, \alpha A + \alpha_S S + \varepsilon_Z) \\
&= Var(Z)\tau + \alpha(\beta_2 - \beta_1) + \alpha_S\beta_S,
\end{aligned}$$

we obtain the regression coefficient of *Z* of the regression of *G* on *Z*,

$b_{GZ} = \tau + \frac{\alpha(\beta_2 - \beta_1)}{Var(Z)} + \frac{\alpha_S \beta_S}{Var(Z)}$ (Equation 5). Since the structural causal model in Figure 1.2A is a restricted model of the structural causal model in Figure 1.2C, we obtain Equation (1.2) by setting $\alpha_S = 0$ and $\beta_S = 0$, that is, $b_{GZ} = \tau + \frac{\alpha(\beta_2 - \beta_1)}{Var(Z)}$.

APPENDIX B: GRAPHICAL REPRESENTATION WITH FLEXIBLE TREATMENT REGIMES

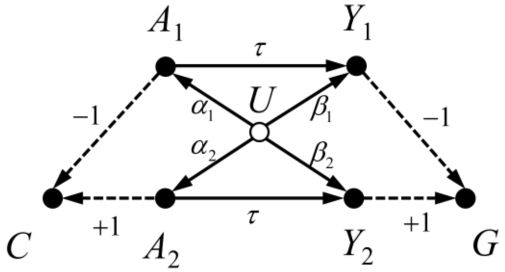

FIGURE A. *Graphical representation, combined the data-generating model (solid arrows) and the analytical model (dashed arrows), for the case where subjects can receive the treatment at both before- and after-stages.*

When subjects can experience the treatment at any time index, the graphical representation can be extended as in Figure A. Instead of a single treatment variable A_2, two different treatment indicators A_1 and A_2, at each time $t=1$ and $t=2$, are displayed. Note that the treatment effect is assumed as being constant across time: Two causal paths $A_1 \rightarrow Y_1$ and $A_2 \rightarrow Y_2$ are denoted by τ. In this case, if the common trend assumption $\beta_1 = \beta_2$ holds, the causal effect can be identified even though the confounder *U* remains unmeasured (a vacant node) by constructing both the treatment difference $C = A_2 - A_1$ and the outcome difference (i.e., gain score) $G = Y_2 - Y_1$.

The treatment difference *C* is associated with the gain score *G* via the following eight paths:

(1) $C \leftarrow A_1 \leftarrow U \rightarrow Y_1 \rightarrow G, \ +\alpha_1\beta_1$;

(2) $C \leftarrow A_1 \leftarrow U \rightarrow Y_2 \rightarrow G, \ -\alpha_1\beta_2$;

(3) $C \leftarrow A_2 \leftarrow U \rightarrow Y_1 \rightarrow G, \ -\alpha_2\beta_1$;

(4) $C \leftarrow A_2 \leftarrow U \rightarrow Y_2 \rightarrow G, \ +\alpha_2\beta_2$;

(5) $C \leftarrow A_1 \leftarrow U \rightarrow A_2 \rightarrow Y_2 \rightarrow G, \ -\alpha_1\alpha_2\tau$;

(6) $C \leftarrow A_2 \leftarrow U \rightarrow A_1 \rightarrow Y_1 \rightarrow G, \ -\alpha_1\alpha_2\tau$;

(7) $C \leftarrow A_1 \rightarrow Y_1 \rightarrow G, \ +\tau Var(A_1)$;

(8) $C \leftarrow A_2 \rightarrow Y_2 \rightarrow G, \ +\tau Var(A_2)$.

The sum of the eight partial associations is given by

$$(\beta_2 - \beta_1)(\alpha_2 - \alpha_1) + \tau\left\{Var(A_1) + Var(A_2) - 2\alpha_1\alpha_2\right\}.$$

By $\beta_1 = \beta_2$ (i.e., common trend) and $Var(C) = Var(A_1) + Var(A_2) - 2\alpha_1\alpha_2$, we see that the covariance between *C* and *G* is $Cov(C, G) = \tau Var(C)$. Therefore, the population regression coefficient of *C* of the regression of *G* on *C* is identical to the causal effect τ:

$$b_{GC} = \frac{Cov(C,G)}{Var(C)} = \frac{\tau Var(C)}{Var(C)} = \tau.$$

Two things are worthy of mentioning. First, the above bias-offsetting mechanism also works when $\alpha_1 = \alpha_2$ holds (even though $\beta_1 \neq \beta_2$). Indeed, this bias-offsetting relying on the equality $\alpha_1 = \alpha_2$ has been known as a negative *treatment* control (instead of a negative *outcome* control) in epidemiology and biomedical research (Lipsitch, Tchetgen Tchetgen, & Cohen, 2010). Our graphical representation reveals the symmetric property of the bias-offsetting mechanism between negative treatment controls and negative outcome controls. Second, the constant treatment effect is crucial in this approach. If $A_1 \rightarrow Y_1$ is denoted by τ_1 and $A_2 \rightarrow Y_2$ is denoted by τ_2 , and $\tau_1 \neq \tau_2$, the resulting the population regression coefficient of *C*, b_{GC} , does not equal either causal effects. In practice, the constant treatment effect across time may be less plausible. One way to circumvent this problem, which is popular in many educational and psychological research, is to design a study so that no subjects are exposed to the treatment at the before-stage. Then, the nodes A_1 and *C* disappear, and researchers investigate the association between A_2 and *G*. This is what we have considered in this paper.

Appendix C: Regression Estimator Formula in Study 2

The population partial regression coefficient for A_2 of the regression model of Y_2 on A_2 and Y_1 can be expressed with bivariate correlations:

$$b_{Y_2A_2|Y_1} = \frac{\rho_{Y_2A_2} - \rho_{Y_2Y_1}\rho_{A_2Y_1}}{1-\rho_{AY_1}^2} \times \frac{SD(Y_2)}{SD(A_2)}.$$

Given $Y_1 = \theta + \varepsilon_1$ and $Y_2 = \tau A_2 + \theta + \varepsilon_2$, the bivariate correlations are given by

$$\rho_{Y_2A_2} = \{Var(A_2)\tau + Cov(A_2,\theta)\} / \{SD(Y_2)SD(A_2)\},$$

$$\rho_{A_2Y_1} = Cov(A_2,\theta) / \{SD(A_2)SD(Y_1)\},$$

$$\rho_{Y_2Y_1} = \{\tau Cov(A_2,\theta) + Var(\theta)\} / \{SD(Y_2)SD(Y_1)\}.$$

Plugging the population correlations into the formula of $b_{Y_2A_2|Y_1}$ above, we obtain Equation (2.5).

Appendix D: Regression Estimator Formulas in Study 3

The structural causal model corresponding to Figure 3.1B is given by $A = \varepsilon_A$, $M = \alpha A + \gamma_M U + \varepsilon_M$, $Y = \tau A + \beta M + \gamma_Y U + \varepsilon_Y$, and $P = \gamma_P U + \varepsilon_P$, where ε_A, ε_M, ε_P, and ε_Y are the mutually independent random disturbance terms (omitted from the graph). Without loss of generality, we assume $Var(U) = 1$. From $Cov(Y, A) = Cov(\tau A + \beta M + \gamma_Y U + \varepsilon_Y, A)$ $= \tau Var(A) + \beta Cov(A, M) = \tau Var(A) + \alpha\beta Var(A)$, Equation (3.1) directly follows. Note that $Cov(A, M) = Cov(A, \alpha A + \gamma_M M + \varepsilon_M) = \alpha Var(A)$.

The partial regression coefficient for *A*, $b_{YA|M}$, of the regression of *Y* on *A* and *M* can be written in terms of bivariate correlations as $b_{YA|M} = \dfrac{\rho_{YA} - \rho_{YM}\rho_{AM}}{1 - \rho_{AM}^2} \times \dfrac{SD(Y)}{SD(A)}$. The correlation coefficients are given by

$$
\begin{aligned}
\rho_{YA} &= Var(A)(\tau + \alpha\beta) / \{SD(Y)SD(A)\}, \\
\rho_{AM} &= Var(A)\alpha / \{SD(A)SD(M)\}, \\
\rho_{YM} &= \{Var(A)\alpha\tau + Var(M)\beta + \gamma_M\gamma_Y\} / \{SD(Y)SD(M)\}.
\end{aligned}
$$

Plugging the population correlations into the formula of $b_{YA|M}$ above, we obtain Equation (3.2).

For Equation (3.3), it is convenient to specify two simple regressions. First, we regress *A* on *P* and obtain the residual $\tilde{A}$ from the regression. Indeed, as $Cov(A, P) = 0$, $\tilde{A} = A$. Second, we regress *M* on *P* and obtain the residual $\tilde{M}$ from the regression. We use $b_{YA|MP} = b_{Y\tilde{A}|\tilde{M}} = b_{YA|\tilde{M}}$. The necessary bivariate correlations among *Y*, *A*, and $\tilde{M}$ are given by

$$
\begin{aligned}
\rho_{A\tilde{M}} &= Var(A)\alpha / \{SD(A)SD(\tilde{M})\}, \\
\rho_{Y\tilde{M}} &= \{Var(A)\alpha\tau + Var(M)\beta + \gamma_M\gamma_Y - \frac{\gamma_M\gamma_P}{Var(P)}(\gamma_P\gamma_Y + \gamma_P\gamma_M\beta)\} / \{SD(Y)SD(\tilde{M})\}.
\end{aligned}
$$

Plugging the population correlations, together with ρ_{YA} above, into the formula of

$b_{YA|\tilde{M}} = \dfrac{\rho_{YA} - \rho_{Y\tilde{M}}\rho_{A\tilde{M}}}{1-\rho^2_{A\tilde{M}}} \times \dfrac{SD(Y)}{SD(A)}$, we obtain Equation (3.3).

In Figure 3.2, the difference score *D* is given by $D = Y - P$. The necessary bivariate correlations to compute the partial regression coefficient $b_{DA|M}$ are

$$\begin{aligned}\rho_{DA} &= Var(A)(\tau+\alpha\beta)/\{SD(D)SD(A)\},\\ \rho_{DM} &= \{Var(A)\alpha\tau + Var(M)\beta + \gamma_M(\gamma_Y - \gamma_P)\}/\{SD(D)SD(M)\}.\end{aligned}$$

Plugging the population correlations, together with ρ_{AM} above, into the formula of

$b_{DA|M} = \dfrac{\rho_{DA} - \rho_{DM}\rho_{AM}}{1-\rho^2_{AM}} \times \dfrac{SD(D)}{SD(A)}$, we obtain Equation (3.5).

Given the structural model corresponding to Figure 3.3, the necessary bivariate correlations are given by

$$\begin{aligned}\rho_{DA} &= \{Var(A)(\tau+\alpha\beta) + \gamma_A\gamma_M\beta + \gamma_A(\gamma_Y - \gamma_P)\}/\{SD(D)SD(A)\},\\ \rho_{AM} &= \{Var(A)\alpha + \gamma_A\gamma_M\}/\{SD(A)SD(M)\},\\ \rho_{DM} &= \{Var(A)\alpha\tau + Var(M)\beta + \gamma_M\gamma_A\tau + \gamma_M(\gamma_Y - \gamma_P)\}/\{SD(D)SD(M)\}.\end{aligned}$$

Plugging the population correlations into the formula of $b_{DA|M}$ above, we obtain Equation (3.6).